## **3D-JIGSAW** (serv) - 65 models for 64 3D targets

### *In Silico* protein recombination

M.N. Offman, S. Wanka, B. Contreras-Moreira*, P.W. Fitzjohn and P.A. Bates
*Cancer Research UK London Research Institute*
paul.bates@cancer.org.uk

Our overall strategy is to enhance protein modelling by considering ensembles of initial models generated from a number of different templates, alignments, scoring functions and algorithms. Our central program module for achieving this is called 'In Silico Protein Recombination' and is based upon a genetic algorithm.

Our methodology is similar for the prediction categories, Comparative Modelling (CM) and Fold Recognition (FR). We have not specifically developed a method for the New Fold (NF) category. Models were collected for most targets from our two fully automatic servers, 3D-JIGSAW-server and 3D-JIGSAW-recomb. Models for all targets for which some human intervention was used were also submitted.

Most methods for one of our servers, 3D-JIGSAW-server, have been described previously[1]. However, the FR module has not been described and can only be briefly outlined here, see below. Methods for our second server, 3D-JIGSAW-recomb, have also been described[2] but there has been some development since CASP5[3], see below.

For all target sequences, and for all methods, the first step is to generate a Position Specific Scoring Matrix (PSSM) and Predicted Secondary Structure (PSS) file. The PSSMs are calculated by PSI-BLAST[4], with five iterations against the nr sequence database (http://www.ncbi.nlm.nih.gov). Each PSS was calculated with PSI-PRED[5] by using the appropriate PSSM described above. These files are then used to score and rank alignments against a library of template PSSM and PSS files for all structural homologues, in the case of potential CM targets, and against a library based on nonredundant PSSM and PSS files (< 30% sequence identity) for potential FR/NF targets. All alignments are generated using the dynamic programming algorithm.

For our FR module seven different functions were used to populate each dynamic programming matrix, these are based upon different PSSM/PSS log-odd mixing ratios – each alignment generated therefore depends upon both target and template PSSM/PSS weighting. For our automatic server, 3D-JIGSAW-server, only the best-ranked alignment is considered further.

For the automatic server, 3D-JIGSAW-recomb, and for all manual submissions, models are constructed with our core 3D-JIGSAW programs using a number of potential templates, plus, alternative alignments to those templates. These are subsequently fed into our genetic recombination algorithm. The steps for this are:

*create initial population of models*
(1) *grow population*: r recombination + (1-r) mutation
(2*) select best proportion according to fitness*
(3) *converged?* stop: otherwise back to (1)

There are some differences in the algorithm compared to that used in CASP5 - new side-chains in every mutation event (1-r) are generated with program SCWRL[6] and recombination events, r, are allowed outside predicted secondary structure elements.

Sometimes, for models involving manual intervention, full three-dimensional models were taken from the CAFASP4 web site and used in the recombination process along with our own models. However, an identical model, to that downloaded from a different server to our own, was never submitted.

1. Bates,P.A. & Sternberg,M.J.E. (1999). Model building by comparison at CASP3: using expert knowledge and computer automation. *Proteins:Struct. Funct. Genet.* **37**, 47-54.
2. Contreras-Moreira,B., Fitzjohn, P.W. & Bates, P.A. (2003). *In silico* protein recombination: enhancing template and sequence alignment selection for comparative protein modelling. *J. Mol. Biol.* **328**, 593-608.
3. Contreras-Moreira,B., Fitzjohn,P.W., Offman,M.N., Smith,G.R. & Bates, P.A. (2003). Novel use of a genetic algorithm for protein structure prediction: searching template and sequence alignment space. *Proteins: Struct. Funct. Genet.* **53**, 424-429.
4. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
5. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* ***292***, 195-202.

6. Dunbrack,R.L.Jr, & Karplus,M. (1993). Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J. Mol. Biol.* **230**, 543-574.

* Did not take part in protein structure prediction for CASP6

## **Accelrys** - 27 models for 16 3D / 1 FN targets

### Homology modeling using a suite of algorithms in Discovery Studio Modeling and Insight II

T. Yeh, J. Fisher-Shaulsky, J. Nauss, Y. Chen,
D. Singh, and D. Haley-Vicente
*Accelrys Inc., 9685 Scranton Rd., San Diego, CA 92121*
dhv@accelrys.com

A plethora of methodologies have been utilized for CASP6 homology model predictions. For ~20 targets within T0196 through T0282, we determined protein models based on a combination of template searching, alignment adjustment, homology modeling, and model refinement and evaluation algorithms available in Discovery Studio® (DS) Modeling and Insight II® modeling and simulations packages (Accelrys, Inc)[1,2].

As part of DS Modeling, an automated, high-throughput functional annotation pipeline program called DS GeneAtlas[3] was used to predict the majority of templates and provide initial alignments and models for each target. The DS GeneAtlas pipeline incorporates sequence similarity detection (e.g. PSI-BLAST), domain analysis (e.g. PFAM), homology modeling (e.g. MODELER), model evaluation (e.g. Profiles-3D), fold recognition (e.g. SeqFold), and 3D active site annotation (e.g. CSC[4] 3D-motif searching) methods. DS GeneAtlas uses a Psi-Blast protocol that combines both direct and reverse search in the profiles space, thus capable of enhancing the homology detection between the query and the template sequences. Using the initial information from DS GeneAtlas, these CASP6 target models were further optimized and evaluated. For a few of the targets, template searching was performed using BLAST, PSI-BLAST and SeqFold. After the template was identified, alignments were adjusted manually or regenerated using Align123 or Align2D (in MODELER). Next, alignments were used for homology modeling using MODELER. These models were then refined (loops and side-chains) using MODELER (Refine_Loop) and Discover. Finally, the models were evaluated using Profiles-3D, MODELER (probably density function values), Prostat/Struct_Check, and Decipher.

Using the CASP6 targets as query sequences, we demonstrate that DS GeneAtlas detects additional relationships, via its high-throughput modeling component, in comparison with the sequence searching method PSI-BLAST only. Furthermore, functionally related proteins with sequence identity below the twilight zone can be recognized correctly. By using a combination of alignment, refinement and evaluation techniques, the best results were achieved for the models.

1. Discovery Studio Modeling (http://www.accelrys.com/dstudio/ds_modeling/) Accelrys Inc.
2. Insight II (http://www.accelrys.com/insight/) Accelrys Inc.
3. Kitson et al. (2002) Functional annotation of proteomic sequences based on consensus of sequence and structural analysis. *Briefings in Bioinformatics* **3**, 1-13.
4. Milik, et al. (2003) Common Structural Cliques: a tool for protein structure and function analysis. *Protein Engineering* **16,** 1-10.

## **Advanced-Onizuka** - 275 models for 64 3D targets

### Fold selection and the SA (GA)-based structure optimization

Kentaro Onizuka
*Advanced Technology Research Laboratories,*
*Matsushita Electric Industrial Co. Ltd.*
onizuka.kentaro@jp.panasonic.com

The method developed to meet CASP6 consists of two units.

1) *Fold recognition unit*
This unit selects hundreds of template conformations that have relatively good compatibility to the target protein sequence among approximately three thousand non-redundant protein structure set collected from PDB. The selected conformations are aligned to the target protein sequence. The compatibility of a conformation to the target sequence is evaluated as the sum of multi-dimensional mean-force potentials between all possible pairs of residues in that conformation (1, 2), now that having the target sequence aligned.

2) *Structure optimization unit*
This unit builds a protein conformation by concatenating the structure segments cut out of those template conformations selected by the fold recognition unit. The templates selected are aligned to the target protein sequence. Here the concatenation of conformations is done as follows; 1) select two (*i*-th and *j*-th respectively) conformations each aligned to the target protein sequence, 2) choose a residue position *M* in the sequence as the crossover point 3) the new conformation is generated by concatenating the segment from N-term (of the target sequence) to *M*-th residue of *i*-th conformation and the segment from *M*-th residue to C-term (of the target sequence) of *j*-th conformation. Since M-th residue is shared by both segments, the relative orientation between the concatenated two segments is fixed. When the generated child conformation has better compatibility to the sequence than both of its parents, the child is selected and the parent having worth compatibility is discarded and is replaced by the child conformation. When the child conformation is slightly worse in compatibility than the parent of better compatibility, there is still a chance for the child to survive. The survival rule follows the Simulated Annealing like criteria with respect to the temperature parameter ; when the temperature is high, the child has big chance to survive, while low temperature, the chance is small. The compatibility to the sequence is the same as that in the Fold recognition unit.

The GA (SA)-based improvements of the conformation are repeated until the conformation converges. The insertion or deletion generated in the alignment process by the Fold recognition unit are, in most cases, automatically swept away during the optimization because those gaps are calculated to have bad score in the compatibility evaluation process..

The performance of the minimization algorithm proposed is intense, although the algorithm logically does not assure to generate the optimal solution.


1. Sippl,M.J. (1990) Calculation of Conformational Ensembles from Potentials of Mean Force: An Approach to the Knowledge-based Prediction of Local Structure in Globular Proteins. *J. Mol. Biol.*, ***213***,859-883.
2. Onizuka,K., Noguchi,T., Akiyama,Y. Matsuda,H. (2002) Using Data Compression for Multidimensional Distribution Analysis. *Intelligent Systems **May/June** 2002*, 48-54.


## **AGAPE-0.3** (serv) - 317 models for 64 3D targets

### AGAPE - fold recognition without template structures


D. Przybylski[1,2,3] and B. Rost[1,2]
*[1] -CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, New York, USA, [2] - Columbia University Center for Computational Biology and Bioinformatics (C2B2), New York, USA,*
*[3] - Depertment of Physics, Columbia University, New York, USA.*
dsp23@columbia.edu


AGAPE[1] is a novel automatic alignment method that uses predicted one-dimensional (1D) structural information (secondary structure and solvent accessibility) for target and template proteins. It is based on an observation that mistakes in the predictions of 1D structure tend to correlate among structurally related proteins. AGAPE uses generalized position specific scoring matrices (sequence + 1D structure) for target and template proteins and a novel 'bi-directional' scoring approach. AGAPE-0.3 is the experimental server under development.


1. Przybylski,D. & Rost,B. (2004). Improving fold recognition without folds. *J Mol Biol* **341**, 255-69.


## **Agata** - 57 models for 52 3D targets

### Modeling of CASP6 target proteins


Agata Chmurzynska
*Agricultural University of Poznan,*
*Department of Animal Genetics and Breeding, Poland*
agata@jay.au.poznan.pl


First step of the procedure was identification of the proteins with known structures related to the targets. Searches with PSI-BLAST [1] were performed against the non-redundant protein database. After inspection with the SWISS PDB Viewer, models were built using SWISS-MODEL program [2].

For more difficult targets, the full protein sequences or their fragments only were submitted to the MetaSever [3]. Selection of the templates was based on the

3D-Jury results [4], and additionally, in some cases alignments were manually modified. Then models were built with the MODELLER program [5].

In order to identify poorly-folded fragments, all the preliminary models were evaluated by Verify 3D [6]. When more than one template was used to create a final model, the initial 3D structures were superimposed and the well-folded fragments were merged. In the final step, missing parts was added using one of the models generated *ab initio* by ROBETTA. This protocol resulted in building several models for almost every target. The final models, submitted to CASP6, were selected after their detailed evaluation with Verify 3D.


1. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W., Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25** (17), 3389-3402.
2. Guex,N., Peitsch,M.C. (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* **18**(15), 2714-2723.
3. Bujnicki,J.M., Elofsson,A., Fischer,D., Rychlewski,L. (2001) Structure prediction meta server. *Bioinformatics*, **17**(8),750-751
4. Ginalski,K., Elofsson,A., Fischer,D., Rychlewski,L. (2003) 3D-jury: a simple approach to improve protein structure predictions. *Bioinformatics* **31**, 3291-3292.
5. Sali,A., Blundell,T.L. (1993) Comparative protein modeling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779-815.
6. Luthy,R., Bowie,J.U., Eisenberg,D. (1999) VERIFY3D: assessment of protein models with three-dimensional profiles. *Nature* **356**, 83-85.


## **ARBY** (serv) - 57 models for 57 3D targets

### The Arby automated structure prediction server


Ingolf Sommer[1], Niklas von Öhsen[2]

*[1] – Max-Planck-Institute for Informatics,*

*[2] – FraunhoferInstitute forScientific Computing and Algorithms*

sommer@mpi-sb.mpg.de


Our fully automated protein structure prediction server Arby[0] combines the results of several fold recognition methods to find suitable templates in a database of structural representatives of protein domains.

The method starts by constructing a set of subsequences from the query sequence, each subsequence representing a hypothesis for a possible protein domain. This is done by scanning against the InterPro database and using hits as domain hypotheses[1]. Additional hypotheses are constructed using a secondary structure prediction from PSIPRED[2]. Segments of predicted loops are used as potential domain boundaries. Finally, the set of subsequences is reduced to a reasonable size by removing subsequences that are highly similar or short.

For each subsequence a multiple alignment is constructed by searching the NR database, clustered to 90% sequence identity, using PSI-BLAST[3]. A frequency profile is calculated from this multiple alignment using a slightly modified version of the Henikoff-Henikoff sequence-weighting algorithm[4].

Each of the potential domains is then subjected to four different fold recognition methods. Each method searches for an optimal structure in our template database. The template database is a representative subset of the SCOP domains with pairwise sequence identity lower than 40% [5, 6]. For each of these template domains, a frequency profile was constructed as described above for the targets. The first fold recognition method is PSI-BLAST, which is used to search through our set of template domains (augmented by the NR sequence database). The second one is the 123D threading program. It uses frequency profiles on the target side and 3D structural information on the template side[7, 8]. The third one is the log-average profile-profile alignment method recently developed in our group[9, 10]. It compares frequency profiles on the target side with profiles on the template side using the log average scoring approach. The fourth method is again the log-average profile-profile alignment program, but in this version it makes use of additional secondary structure information on the target and template side.

The quality of each of these search results is assessed using confidence measures. For PSI-BLAST, these are readily available[11], for the other methods, we use empirical confidence measures[12].

The target sequence is then annotated with all the produced quadruplets (subsequence, fold recognition method, search result, confidence value). Finally, we select a set of non-overlapping annotations along the sequence, by performing greedy optimization on the confidence values. For each of these selected annotations, a separate protein domain is predicted. The structure of this domain prediction is computed by aligning the subsequence against the template structure using log-average profile-profile alignment.

The underlying machinery is a Java based data flow engine, designed for stability. Since it is general and independent of the specific pipeline (as the one described above), it can be used as infrastructure for other projects as well: we developed a component framework in which all algorithms and programs are encapsulated in small Java classes. Each of these components specifies an algorithm to be executed along with its input parameters, the output that it produces, and possible error conditions. The accompanying engine provides a number of features for the components: First of all, the input/output dependencies of components are resolved. If all inputs for a specific algorithm have been determined, the algorithm itself is being scheduled for execution. The components are executed in parallel on any number of CPUs, in our case 64 CPUs of a SunFire 15000 server. A frequent problem in fully automated systems is reliable error handling. We solve this problem by catching potential error conditions and adaptively pruning the data-flow tree.

In a nutshell, the structure prediction server is based on the use of profile-profile algorithms for fold recognition, the quality assessment using confidence measures, and the stable and powerful Java data flow engine

Acknowledgements

In addition to the authors, the ARBY Team includes Mario Albrecht, Thomas Lengauer, Theo Mevissen, Oliver Sander, and Ralf Zimmer. We thank Daniel Hanisch for providing contributions to the Java implementation. Part of this research has been supported by BMBF grant no. 01 SF 9984/3 (Helmholtz Network for Bioinformatics) and DFG grant no. Le 491/14.

0. von Öhsen,N., et al. (2004) Arby: automatic protein structure prediction using profile-profile alignment and confidence measures. *Bioinformatics*. **20** (14), 2228-35.
1. Apweiler,R., et al. (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res*. **29** (1), 37-40.
2. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*. **292** (2), 195-202.
3. Altschul,S.F., et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. **25** (17), 3389-402.
4. Henikoff,S. and Henikoff,J.G. (1994) Position-based sequence weights. *J Mol Biol*. **243** (4), 574-8.
5. Chandonia,J.M., et al. (2002) ASTRAL compendium enhancements. *Nucleic Acids Res*. **30** (1), 260-3.
6. Brenner,S.E., Koehl,P., and Levitt,M. (2000) The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res*. **28** (1), 254-6.
7. Zien,A., Zimmer,R., and Lengauer,T. (2000) A simple iterative approach to parameter optimization. *J Comput Biol*. **7** (3-4), 483-501.
8. Alexandrov,N.N., Nussinov,R., and Zimmer,R. (1996) Fast protein fold recognition via sequence to structure alignment and contact capacity potentials. *Pac Symp Biocomput*, 53-72.
9. Von Öhsen,N., Sommer,I., and Zimmer,R. (2003) Profile-Profile Alignment: A Powerful Tool For Protein Structure Prediction. in *Pac Symp Biocomput*.
10. Von Öhsen,N. and Zimmer,R. (2001) Improving profile-profile alignment via log average scoring. *Lecture Notes in Computer Science*. **2149**, 11-26.
11. Karlin,S. and Altschul,S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci U S A*. **87** (6), 2264-8.
12. Sommer,I., et al. (2002) Confidence measures for protein fold recognition. *Bioinformatics*. **18** (6), 802-12.

**Atid** - 17 models for 17 3D targets

## Structure prediction through direct folding simulation

J. Rosenzweig[1] and I. Rosenzweig[2]
*[1] – Cambridge Proteomics Ltd, [2] – Addenbrooke's Hospital*
j.rosenzweig@camprot.com

Protein structures are predicted via direct simulation of protein folding pathways.

Starting from extended initial configurations, proteins were followed through their folding pathways to compact, folded free energy minima at temperatures slightly exceeding 300K.

Force fields used were CHARMM17 (united atom) and AMBER99 (all atom), and the solvation energies were computed using GBSA-type models.

All simulations were performed using the novel *in-house* ATID protocol, on two dual Linux workstations running at 2x2.1 GHz and 2x2.8 GHz, respectively.

Each folding simulation was run twice, and only the results in which both structures agreed to within 1A rmsd were submitted in order to eliminate the possibility of chaperone-assisted folding.

## **B213-207** - 320 models for 64 3D / 1 FN targets

### Optimization of predicted spatial restraints on a coarse-grained protein model

O. Venezuela[1], Y.H. Tan[1] and D. Kihara[2,1]
[1] – *Dept. of Computer Science,* [2] – *Dept. of Biological Sciences, Purdue University, West Lafayette, IN, USA*
dkihara@purdue.edu

It has been shown that in many cases the recent generation of fold recognition methods can capture at least structural fragment information even when the global structure can not be reliably predicted[1]. Assembling structure fragments detected by a fold recognition method is one of the common ways for *ab initio/de novo* protein structure prediction[1-3]. In the CASP5, it was reported that several consensus methods or meta-server approaches[4,5] showed high performance[6]. Based on these two observations, our approach developed for CASP6 is an optimization of predicted spatial restraints calculated by various servers, including servers participating in CAFASP4 using a coarse-grained protein model.

A protein is represented by a simplified model which explicitly specifies positions of alpha carbons in the main chain[7]. A conformation of this Cα model is defined by a set of rotational and hinge angles between adjacent alpha carbons. Information of predicted structures of a target protein by various methods is used in the following way: (1) The predicted structures are clustered globally and locally. (2) The distribution of inter-residue distances and angles are calculated and subsequently used as soft restraints[8] in the next refinement step. Starting from several initial structures, the conformation of the model is refined so that is satisfies these soft restraints by a Monte Carlo optimization with the Metropolis criteria. Consensus prediction of the secondary structures is also used. Usually a conformation converges relatively quickly since the method uses a large number of restraints.

During the course of the development, we phased in statistics of the structure preference of known structures in PDB as penalty terms of spatial restraints to avoid "non-protein-like" conformations[7]. These terms include minimum distance between Cα - Cα , peptide bond - peptide bond, and Cα - peptide bond distances as well as hinge angle restraints.

Suggestions from our function prediction team were often a great help in the final model selection. Our three teams, the structure, function (B213-207Func), and domain prediction (B213-207Dom) teams worked in a coordinated manner. Although this method is still in an early stage of the development, the performance will surely improve as additional scoring terms are incorporated.

1. Kihara,D., Lu,H., Kolinski,A. & Skolnick,J. (2001) TOUCHSTONE: an ab initio protein structure prediction method that uses threading-based tertiary restraints. *Proc Natl Acad Sci U S A* **98**, 10125-30.
2. Jones,D.T. (2001) Predicting novel protein folds by using FRAGFOLD. *Proteins* **Suppl 5**, 127-132.
3. Bonneau,R. *et al.* (2002) De Novo Prediction of Three-dimensional Structures for Major Protein Families. *J Mol Biol* **322**, 65.
4. Fischer,D. (2003) 3D-SHOTGUN: A novel, cooperative, fold-recognition meta-predictor. *Proteins* **51**, 434-441.
5. Ginalski,K. & Rychlewski,L. (2003) Detection of reliable and unexpected protein fold predictions using 3D-Jury. *Nucleic Acids Res.* **31**, 3291-3292.
6. Kinch,L.N. *et al.* (2003) CASP5 assessment of fold recognition target predictions. *Proteins* **53 Suppl 6**, 395-409.
7. Kolinski,A. (2004) Protein modeling and structure prediction with a reduced representation. *Acta Biochim. Pol.* **51**, 349-371.
8. Sali,A. & Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **234**, 779-815.

## **B213-207Func** - 68 models for 64 FN targets

### A structured approach to computational function prediction

T. Hawkins[1] and D. Kihara[1,2]
[1] – *Dept. of Biological Sciences, Purdue University,* [2] – *Dept. of Computer Science, Purdue University, West Lafayette, IN, USA*
dkihara@purdue.edu

For function prediction in CASP6, we used a multi-layered, multi-dimensional approach. The process of defining functions for uncharacterized protein targets involved three steps: (1) searching the primary target sequence against functional databases, (2) manually building and refining data from primary searches, and (3) assigning GO numbered definitions to predicted functions. This method was used to gather predictions for the GO Molecular Function, GO Biological Process, and GO Cellular Component categories. BLAST and PSI-BLAST[1] were used for sequence similarity; PROSITE[2], PRINTS[3] and

Blocks[4] were used for functional motif searching; Pfam and Pfam-FS[5] were used to for family alignments; PSORT[6] was used for subcellular localization; and STRING[7] was used for additional functional associations in primary searches. Information in the KEGG Pathway database[8] and thorough literature searches were used refine and build on the data gathered from primary searches in the cases where that data was not sufficient to make a reasonable prediction of GO categories. GoFigure[9] and AmiGO[10] were used to find GO definitions for predicted functions.

To predict binding sites, multiple sequence alignments were made using ClustalW of BLAST and PSI-BLAST hits below an e-value of 0.01 (limited to 20). Conserved regions were determined manually and localized on predicted structures; regions containing clusters of conserved residues were predicted to be binding sites. If the predicted function of the protein indicated binding of a specific partner, that molecule/macromolecule was predicted to interact with the predicted binding region. If a conserved region consisted of 5 or more consecutive residues, we considered it to be a functional motif. All of these motifs for a single target sequence were searched individually against the NR protein database in the cases where other data was not sufficient to make a reasonable prediction.

Using this method, reasonable predictions were made for each of the 76 valid protein targets in CASP6. Automation of this method, including substitution of rule-based algorithms for manual interpretation steps, is underway in preparation for function prediction in CASP7.


1. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402. [http://www.ncbi.nih.gov/BLAST/].
2. Sigrist,C.J.A., Cerutti,L., Hulo,N., Gattiker,A., Falquet,L., Pagni,M., Bairoch,A. & Bucher,P. (2002). PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform*. **3**, 265-274. [http://www.expasy.org/prosite/].
3. Attwood,T.K., Bradley,P., Flower,D.R., Gaulton,A., Maudling,N., Mitchell,A.L., Moulton,G., Nordle,A., Paine,K., Taylor,P., Uddin,A. & Zygouri,C. (2003). PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res.* **31**, 400-402. [http://bioinf.man.ac.uk/dbbrowser/PRINTS/].
4. Henikoff,S., Henikoff,J.G. & Pietrokovski,S. (1999). Blocks+: A non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics.* **15**, 471-479. [http://blocks.fhcrc.org/].
5. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L.L., Studholme,D.J., Yeats,C. & Eddy,S.R. (2004). The Pfam Protein Families Database. *Nucleic Acids Res.* **32**, D138-D141. [http://www.sanger.ac.uk/Software/Pfam/].
6. Nakai,K. & Kanehisa,M. (1991). Expert system for predicting protein localization sites in Gram-negative bacteria. *PROTEINS: Structure, Function, and Genetics*. **11**, 95-110. [http://psort.nibb.ac.jp/].
7. von Mering,C., Huynen,M., Jaeggi,D., Schmidt,S., Bork,P. & Snel,B. (2003). STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res*. **31**, 258-261. [http://string.embl.de/].
8. Kanehisa,M. & Goto,S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27-30. [http://www.genome.jp/kegg/pathway.html].
9. Khan,S., Situ,G., Decker,K. & Schmidt,C.J. (2003). GoFigure: automated Gene Ontology annotation. *Bioinformatics*. **19**, 2484-2485. [http://udgenome.ags.udel.edu/gofigure/].
10. AmiGO. [http://www.godatabase.org/].


## **BAKER** - 433 models for 64 3D / 63 RR / 58 FN targets

### Novel approaches to protein structure prediction at CASP6


P. Bradley, G. Cheng, D. Chivian, D. Kim, L. Malmstrom, J. Meiler, K. Misura, Bin Qian, J. Schonbrun, A. Zanghellini, D. Baker*

*University of Washington*

dabaker@u.washington.edu


<u>*Domain Parsing (DK).*</u> Targets were parsed into putative domains based on the results obtained from GINZU and ROSETTADOM. For more difficult targets, alternative domain boundaries were considered and final models were chosen through manual human inspection.

Targets with 3D JURY[1] A1 score of 50 or larger, or with PDB homology hits of e-value lower than 0.001 as defined by PSIBLAST[2] are considered as fold recognition or comparative modeling targets, respectively. All remaining targets are modeled *de novo*.

<u>*De novo prediction (PB, LM & KM).*</u> The protocol for *de novo* prediction focused on generating diverse populations of structural models. Diversity was achieved by folding large numbers of sequence homologs; by generating

decoys in multiple rounds using modified parameters (mainly secondary structure predictions) for later generations; and by post-filtering of large decoy ensembles to explore under-sampled topologies. As in previous CASP's, decoys were generated using the ROSETTA[3] fragment assembly algorithm.

In addition, we tested a new protocol for generating models with long-range beta-sheet pairings: given a set of target pairings, models are constructed in which the paired residues are maintained in a constant relative orientation corresponding to ideal β-sheet geometry. A number of chain breaks equal to the number of long-range constraints are introduced to ensure that a unique structure can be constructed from the torsion angles, and structures are generated and scored by fragment assembly as in standard ROSETTA. Target pairings can be chosen to sample a range of non-local topologies, or by analysis of frequently sampled pairings in the fold-recognition server results (Error: Reference source not found).

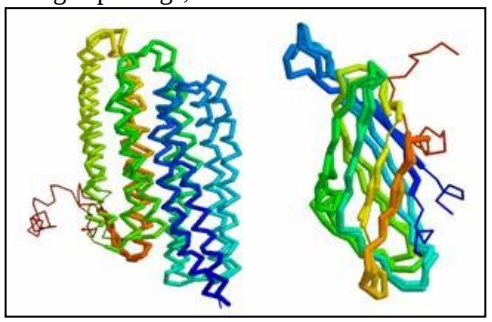


*Figure 0: Target structure and de novo models for T198 (left) and T212 (right). The model for T212 was built using the novel approach for non-local β-sheet contacts.*

Final model selection was generally based on clustering of the decoys. All-atom models were built for some of the smaller targets and ROSETTA's high-resolution refinement protocol and scoring function were used to select submissions.

*Comparative Modeling and Fold Recognition (BQ & DC).* Parent detection is performed by the ROBETTA04 protocol (see accompanying abstract), where BLAST/PSIBLAST, FFAS03, or 3D JURY scores are used to select the parents with highest confidence. When there are multiple parents with similar confidence scores, all distinct parents are used in the subsequent modeling process.

We use target-parent alignments from several different sources: 1) ROBETTA04 server alignments, which are selected by physical energies of structure models that were built based on an alignment ensemble. 2), 3D JURY server alignments, which are selected using the consensus alignments from different alignment methods. 3), Manual alignments based on PSIBLAST sequence profiles, aided by functional information from literature search. These alignments are compared and the representatives are used to model the aligned regions of the targets.

The structural core regions of the targets that have hits in PDB with PSIBLAST e-values of 0.001 or lower are allowed to be flexible and refined using a physical energy based refinement protocol. In this protocol, the principle components of the variation observed in structural homologs are used to define the preferred backbone conformational space. A grid sampling in this preferred conformational space generates a structural model ensemble, which is subject to Rosetta full-atom energy evaluation. The models with the lowest physical energies are selected for further modeling of the loop regions.

*Loop Modeling (JS).* We employed a novel atomic resolution procedure to model unaligned segments in homology models. For loops under 17 residues we searched through the Protein Databank to find a large population (~2000) of segments with good profile-profile matches to the target in the loop region, and where the distance between then Cα atoms at each end of the loop were close to the distance in the parent structure. For longer loops we used the standard ROSETTA *de novo* fragment insertion method to generate an initial population of loops. All loops were then closed using the analytic Cyclic Coordinate Decent method[4]. Finally, side-chains were added, and all loop regions were refined using our atomic resolution potential. The five loops with the lowest energies were selected for CASP6. This method was used for the unaligned regions of BLAST and PSIBLAST[2] detectable homologs.

*Domain Assembly (AZ).* For multi-domain targets, a specific mode of ROSETTA is used to assemble the individual models generated either from fold recognition, homology modeling or *de novo* folding. Once the linker region is defined, fragments are inserted exclusively into the linker region using a MONTE CARLO procedure analogous to the one used in *de novo*. After each insertion the total energy is computed. Subsequently the side-chain centroid decoys are clustered. The cluster centers and the 100 lowest scoring structure undergo a refinement with the atomic resolution potential using a sequence of small moves in the linker region. After each move the structure is repacked and its score is evaluated.

*Consensus Contact Prediction (JM).* Based on the protein structure predictions of 24 servers that participated in the LIVEBENCH 7 and LIVEBENCH 8 experiments[5] (357 targets in total) an artificial neural network was trained to perform a consensus contact prediction.

The network is setup to predict a potential contact between two amino acids. By sweeping over all pairs of amino acids the whole contact map can be predicted. All amino acid pairs having their Cα atoms closer than 11Å were considered as

being in contact if they are separated by more than at least 8 more amino acids in sequence in order to focus on non-local contacts.

Input to the neural network are position of the amino acids in sequence, JUFO secondary structure prediction (*www.jens-meiler.de/jufo.html*), as well as position specific scoring matrices from PSIBLAST[2] for two windows of 5 amino acids around the amino acids of interest. In addition the contacts predicted in the top five models of the 24 servers are used together with the respective scores as input. The output range is [0,1] with 0 being no contact and 1 being contact.

The results are summarized in Table 1. At an output level of 0.7 the network predicts approximately half the contacts correctly by mis-predicting only 3% of the non-contacts as contacts.

Table 1: ANN results if output levels above 0.5, 0.7, or 0.9 are counted as predicted contact.

| output level: | 0.5 | | 0.7 | | 0.9 | |
|---|---|---|---|---|---|---|
| ANN prediction | Target contacts (left) and non-contacts (right) | | | | | |
| contacts: | 70% | 13% | 49% | 3% | 19% | 0% |
| non-contacts: | 30% | 87% | 51% | 97% | 81% | 100% |

*Function Prediction (GC).* For *de novo* targets, we used a motif based algorithm to search decoy ensembles to identify the potential function of the target. For other targets GO annotations from fold matches was utilized. Predictions for functional or binding sites on the BLAST/PSIBLAST level rely on homology based binding site mapping. Ligands from the template PDB were mapped onto corresponding regions of the model. For FR level targets, the model and template are chosen based on functional site conservation. For *de novo* targets, the function site is mainly predicted by sequence conservation.


1. Ginalski,K., Elofsson,A., Fischer,D. & Rychelewski,L. (2003). 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics* **19**, 1015-1018.
2. Altschul,S.F., Madden, T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
3. Simons,K.T., Kooperberg,C., Huang,E. & Baker,D. (1997). Assembly of Protein Tertiary Structures from Fragments with Similar Local Sequences using Simulated Annealing and Bayesian Scoring Functions. *J. Mol. Biol.* **268**, 209-225.
4. Canutescu,A. A. & Dunbrack,R.L. (2003). Cyclic coordinate descent: A robotics algorithm forprotein loop closure. *Protein Sci* **12**, 963-972.
5. Bujnicki,J.M., Elofsson,A., Fischer,D. & Rychlewski,L. (2001). LiveBench-2: Large-Scale Automated Evaluation of Protein Structure Prediction Servers. *Proteins: Struct., Funct., Genet.* **Suppl5**, 184-195.


## **BAKER-ROBETTA** (serv) - 320 models for 64 3D targets
## **BAKER-ROBETTA_04** - 320 models for 64 3D targets

### The Robetta and Robetta_04 protocols


Dylan Chivian[1], David E. Kim[1], Lars Malmstrom[1],
Jack Schonbrun[1], Carol A. Rohl[2] & David Baker[1,*]
*1- University of Washington, Seattle, WA*
*2- University of California, Santa Cruz, CA*
dabaker@u.washington.edu


The Rosetta[1] homology modeling and *de novo* protocols for protein domain prediction have been incorporated with the Ginzu homolog identification and domain parsing protocol into an automated protocol called Robetta[2,3] to allow for tertiary structure prediction for the full length of a protein chain. We have modified the Robetta homology modeling protocol from that used in CASP-5 to include energetic selection from a model ensemble. Additionally, in the Robetta_04 homology modeling protocol, we investigate the effectiveness of modeling based on multiple parents, loop optimization, and global optimization for fold recognition targets. The Robetta_04 *de novo* protocol investigates the effect of re-ranking models based on a confidence score. Robetta, which participated in CASP as a server, is fully automated, and currently offered as a server to the public at http://robetta.bakerlab.org/. Robetta_04, due to the lack of complete automation participated in CASP as a non-server group, is nonetheless mostly automated. The remainder of the protocol is followed closely and without application of human intuition with the intent of future inclusion of successful ideas into the fully automated server, as well as to serve as a control to compare with our human group's results.

Robetta homology modeling protocol

Robetta uses the highest confidence detection (or the longest detection if similar in confidence) from BLAST/PSI-BLAST[4], FFAS03[5], or 3DJury-A1[6] to select the parent for homology modeling. Important to note is that Robetta does not use the alignment from the detection method except to determine the

domain(s) of the parent to model against. Rather it parametrically generates its own alignment ensemble using the K*Sync alignment method[2] by varying the sequence profile comparison method, the source of the secondary structure prediction, the stringency of the sequence profile, the stringency of the StrAD-Stack multiple structural alignment used to define obligate elements, and the weights on the terms in the dynamic programming scoring function. The alignment ensemble is turned into a decoy ensemble by placing the sequence of the query onto the backbone of the parent based on the alignment. Unaligned loop regions are assembled from fragments and optimized to fit the aligned template structure[7]. Side-chains are added using a backbone-dependent rotamer library[8] with a Monte Carlo conformational search procedure[9]. The template region is kept fixed, and models are selected from the ensemble using variants of the Rosetta energy function.

Robetta *de novo* protocol

Robetta *de novo* modeling generates 10000 query decoys and 5000 decoys for up to 2 homologous sequences using the Rosetta fragment-assembly methodology[10]. Those decoys are filtered down to 2000 for the query and 1000 for each homolog in order to down-weight Rosetta pathologies, such as low contact-order structures. The filtered ensemble is structurally clustered, and the top 5 cluster centers by population are returned in order as the predictions.

Robetta_04 homology modeling protocol

Robetta_04, like Robetta, examines ensembles of alignments produced parametrically with the K*Sync alignment method, but includes up to 5 parents. Loops are optimized for closure[11] and energy with the template. PSI-BLAST level targets have frozen templates plus loops modeled by fragments, with models selected from the ensemble by the Rosetta full-atom energy function. Targets in the fold recognition category, those detected by FFAS03 and 3DJury, are allowed backbone flexibility along the entire chain, including template regions, during optimization of Rosetta's side-chain centroid energy function with fragment-insertion. Final predictions are selected from the optimized ensemble by the Rosetta side-chain centroid energy function.

Robetta_04 *de novo* protocol

The only difference in the Robetta_04 *de novo* protocol from the Robetta *de novo* protocol is in the approach for selecting the final predictions. Rather than return cluster centers in order by population, the cluster centers are scanned against known PDB structures with MAMMOTH[12]. A confidence function, similar to one used previously[10], incorporating the significance of any MAMMOTH hit, the length of the MAMMOTH match, the contact order of the decoy, and the clustering convergence, is used to re-rank the cluster centers to determine the top models.


1. Simons,K.T., Kooperberg,C., Huang,E., & Baker D. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol* **268**, 209-225.
2. Chivian,D., Kim,D.E., Malmstrom,L., Bradley,P., Robertson,T., Murphy,P., Strauss,C.E., Bonneau,R., Rohl,C.A., & Baker,D. (2003) Automated prediction of CASP-5 structures using the Robetta server. *Proteins* **53**, 524-533.
3. Kim,D.E., Chivian,D., & Baker,D. (2004) Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res* **32**, W526-W531.
4. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W., & Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402.
5. Rychlewski,L., Jaroszewski,L., Li,W., & Godzik,A. (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci* **9**, 232-241.
6. Ginalski,K., Elofsson,A., Fischer,D., & Rychlewski,L.. (2003) 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics* **19**, 1015-1018.
7. Rohl,C.A., Strauss,C.E., Chivian,D., & Baker,D. (2004) Modeling structurally variable regions in homologous proteins with rosetta. *Proteins* **55**, 656-677.
8. Dunbrack,R.L., & Cohen,F.E. (1997) Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci* **6**,1661-1681.
9. Kuhlman,B., & Baker,D. (2000) Native protein sequences are close to optimal for their structures. *Proc Natl Acad Sci USA* **97**, 10383-10388.
10. Bonneau,R., Strauss,C.E., Rohl,C.A., Chivian,D., Bradley,P., Malmstrom,L., Robertson,T., & Baker,D. (2002) De novo prediction of three-dimensional structures for major protein families. *J Mol Biol* **322**, 65-78.
11. Canutescu,A.A., & Dunbrack,R.L. Jr. (2003) Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Sci* **12** 963-972.
12. Ortiz,A.R., Strauss,C.E., & Olmea,O. (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci* **11**, 2606-2621.

## **BAKER-ROBETTA-GINZU** (serv) 64 models for 64 DP targets

### The Ginzu homologue identification and domain parsing protocol

Dylan Chivian, David E. Kim, Lars Malmstrom, & David Baker*
*University of Washington, Seattle, WA*
dabaker@u.washington.edu

Protein chains often contain more than one domain. In order to predict the domain organization of a protein, we have developed the Ginzu homolog identification and domain parsing method and applied it in the Robetta server[1,2] to allow for domain-based tertiary structure prediction of the full length of protein chains. The method is available to the public as part of the Robetta server (http://robetta.bakerlab.org/).

Ginzu attempts to determine the locations of putative domains in the query sequence and the identification of any likely homologs with experimentally characterized structures. These steps are not decoupled, since the ability to assign a region of the target to a known protein structure greatly increases the likelihood that it is at least one protein domain. The approach consists of scanning the target sequence with successively less confident methods to assign regions that are likely to be domains. Once those regions are identified, cut points in the putative linkers are determined, and if possible a single homologous PDB chain is associated with each putative domain. The initial scan attempts to identify the closest relatives with experimental structures to regions of the query sequence. A straightforward BLAST[3] search against the PDB sequence database detects such relatives. All PDB ids that are detected at this stage are stored. A PSI-BLAST[3] search is then used to detect more distant relatives of the query, as well as provide more complete coverage since such alignments tend to be longer. Non-overlapping regions that possess the best combination of detection confidence and length of coverage are assigned as domains. The associated PDB id and region of the chain matched is retained.

One may then employ more remote fold-recognition methods to detect homologous PDB structures. We used FFAS03[4] in this step for the parsing of the CASP-6 targets. Again, as with the PSI-BLAST detections, the associated PDB and region of the target chain covered is retained. Following the FFAS03 step, we scan remaining regions with 3D-Jury-A1[5] in the same fashion. Detected fold relatives with structures are stored.

Any remaining long regions of the query that do not have structural homologs identified may require further division into putative domains. After all regions of the query that are likely a contiguous domain (or domains) based on homology to a PDB structure have been assigned, one may continue to determine regions that have increased likelihood of being a single domain by applying a HMMER search of Pfam[6]. Subsequent steps of Ginzu utilize the program "msa2domains", which examines the PSI-BLAST multiple sequence alignment (MSA) to find clusters of sequences in the PSI-BLAST multiple sequence alignment (MSA) and assigns these as regions of increased domain confidence for any stretches of the target that have not yet been found to have a domain. This is done in an order based on the number of unique observations in the cluster (essentially a non-redundant depth), with overlaps not permitted. Lastly, msa2domains determines where to place the exact cut points in the linker regions, or any remaining long unassigned regions, via a heuristic that again considers clusters of sequences in the PSI-BLAST MSA, the least occupied positions in the MSA, strongly predicted loop regions by PSIPRED[7], and distance from the nearest region of increased domain confidence. A fourth term boosts the likelihood of a domain boundary in regions of the MSA where the sequences frequently begin or end.

The final step consists of parsing regions that have been assigned structural homologs based on the model generated by that assignment. We have developed a consensus variant of Taylor's structure-based domain parsing method[8] that is applied to the target model as well as PSI-BLAST detectable structural homologs to complete the domain parsing.

1. Chivian,D., KimD.E., Malmstrom,L., Bradley,P., Robertson,T., Murphy,P., Strauss,C.E., Bonneau,R., Rohl,C.A., & Baker,D. (2003) Automated prediction of CASP-5 structures using the Robetta server. *Proteins* **53**, 524-533.
2. Kim,D.E., Chivian,D., & Baker,D. (2004) Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res* **32**, W526-W531.
3. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W., & Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402.
4. Rychlewski,L., Jaroszewski,L., Li,W., & Godzik,A. (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci.* **9**, 232-241.
5. Ginalski,K., Elofsson,A., Fischer,D., & Rychlewski,L. (2003) 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics* **19**, 1015-1018.
6. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R.., Griffiths-Jones,S., Howe,K.L., Marshall,M., & Sonnhammer,E.L. (2002)

The Pfam protein families database. *Nucleic Acids Res.* **30**, 276-280.
7. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol. Biol.* **292**, 195-202.
8. Taylor,W.R. (1999) Protein structural domain identification. *Protein Eng.* **12**, 203-216.

## **BAKER-ROSETTADOM** (serv) - 64 models for 64 DP targets

### The RosettaDOM domain parsing protocol

D.E. Kim, D. Chivian, L. Malmstrom and D. Baker
*University of Washington*
dabaker@u.washington.edu

Predicting protein domain boundaries accurately is a difficult yet important step in protein structure prediction. Here, we describe a protocol to identify protein domain boundaries using a sequence homology based procedure called Ginzu[1-2], and an ab initio method that uses the Rosetta[3-5] structure prediction software suite for proteins lacking significant homology to experimentally determined structures.

RosettaDOM first uses Ginzu to identify domains that are homologous to known structures in the PDB. See accompanying Ginzu abstract for details. If Ginzu assigns a domain based on homology to a known structure in the PDB using either BLAST[6], PSI-BLAST[6], or FFAS03[7], RosettaDOM simply returns the domain boundary predictions provided by Ginzu. For query sequences lacking such homology, an ab initio domain prediction method similar to SnapDRAGON[8] is used. The ab initio method consists of generating 400 three-dimensional models using Rosetta, and then selecting 200 models based on score and whether they pass filters that eliminate structures with too many local contacts or unlikely strand topologies. Domain boundaries are then assigned for each of the 200 models using a structure based domain identification algorithm[9]. Final domain boundary predictions are made based on consistencies found in the domain assignments of these models. Domain boundaries are chosen under the assumption that although Rosetta is unlikely to produce accurate atomic-resolution models, it may accurately produce coarse structural features such as domains. An example of this was shown for T148 in CASP5[1;3].

1. Chivian,D., Kim,D.E., Malmstrom,L., Bradley,P., Robertson,T., Murphy,P., Strauss,C.E., Bonneau,R., Rohl,C.A. & Baker,D. (2003). Automated prediction of CASP-5 structures using the Robetta server. *Proteins* **53 Suppl 6**, 524-533.
2. Kim,D.E., Chivian,D. & Baker,D. (2004). Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res.* **32(Web Server issue)**, W526-W531.
3. Bradley,P., Chivian,D., Meiler,J., Misura,K.M., Rohl,C.A., Schief,W.R., Wedemeyer,W.J., Schueler-Furman,O., Murphy,P., Schonbrun,J., Strauss,C.E. & Baker,D. (2003). Rosetta predictions in CASP5: successes, failures, and prospects for complete automation. *Proteins* **53 Suppl 6**, 457-468.
4. Bonneau,R., Strauss,C.E., Rohl,C.A., Chivian,D., Bradley,P., Malmstrom,L., Robertson,T. & Baker,D. (2002). De novo prediction of three-dimensional structures for major protein families. *J. Mol. Biol.* **322**, 65-78.
5. Simons,K.T., Ruczinski,I., Kooperberg,C., Fox,B., Bystroff,C., & Baker,D. (1999). Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins* **34**, 82-95.
6. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W., & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. **25**, 3389-3402.
7. Rychlewski,L., Jaroszewski,L., Li,W., & Godzik,A. (2000). Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci*. **9**, 232-241.
8. George,R.A. & Heringa,J. (2002). SnapDRAGON: a method to delineate protein structural domains from sequence data. *J. Mol. Biol*. **316**, 839-851.
9. Taylor,W.R. (1999). Protein structural domain identification. *Protein Eng*. **12**, 203-216.

## **Baldi-group** - 434 models for 64 3D / 9 DP / 23 DR / 23 RR targ

### Prediction of protein structural features and tertiary structures from SCRATCH using recursive neural networks, evolutionary information, fragment libraries, and energy functions

P.F. Baldi, J. Cheng, A.Z. Randall, M. J. Sweredoski
*University of California, Irvine*
pfbaldi@ics.uci.edu

Our CASP predictions of protein domains, disordered regions, contact maps, and 3D structures are based on the latest version of our SCRATCH suite of predictors. The suite combines machine learning methods, evolutionary information in the form of profiles, fragment libraries extracted from the PDB[1], and energy functions to predict protein structural features and complete structures. The suite includes the following main modules:

- SSpro[2]: secondary structure
- ACCpro[3]: relative solvent accessibility
- MUpro: effect of single AA mutation on stability
- DISpro: disordered regions
- DOMpro: domains
- DIpro: disulphide bridges
- CMAPpro[4,5]: contact maps at 6, 8, 10, and 12 Å
- CCMAPpro : coarse contact maps
- 3Dpro: 3D structure

All predictors are periodically trained in a supervised fashion and cross-validated using curated, non-redundant, datasets extracted from the PDB. Structural feature predictors (SSpro, ACCpro, MUpro, DISpro, and DOMpro) use ensembles of 1D-RNN (one dimensional- recursive neural network) architectures[5]. Contact map (CMAPpro and CCMAPpro) and disulphide bridge (DIpro) predictors use ensembles of 2D-RNN architectures[4,5] ([DIpro also uses kernel methods]. These architectures are based on probabilistic graphical models (Bayesian networks) meshed with a neural network parameterization to accelerate belief propagation and learning. These architectures systematically combine standard information contained in a local input window with more distant contextual information extracted by translation-invariant recursive neural networks that are convolved along the entire length of the protein (1D) or of the contact maps (2D) from all possible directions.

All predictors, except 3Dpro, directly leverage homology information in the form of input profiles derived using PSI-BLAST[6] to include remote homologs[7,8]. In addition, very high-levels of local homology to known structures are used either directly or in combination with the output of the corresponding predictors. For instance, the secondary structure and solvent accessibility of homologous fragments are combined with the outputs of SSpro and ACCpro to improve their prediction accuracy for target sequences. Whenever possible and useful, predictors leverage the output of the other predictors and use them as part of their inputs. For instance, the outputs from SSpro (secondary structure) and ACCpro (solvent accessibility) are fed into DOMpro for domain boundary prediction and into 3Dpro for tertiary structure prediction.

Taking sequence profile, predicted secondary structure, and solvent accessibility as inputs, DISpro predicts the disordered/ordered state for each residue in the sequence using ensembles of 1D-RNNs. DOMpro produces domain prediction in three steps. First, using the same inputs as DISpro and the same recursive neural network architectures, DOMpro predicts whether a residue belongs to a domain boundary region or not. Residues within 20 amino acids from the actual domain boundary as annotated in the CATH[9,10] database are considered to be part of the domain boundary region. Second, a statistical approach is used to infer the domain boundaries from the predicted states (boundary/non-boundary) of the individual residues. Finally, the sequence segments separated by domain boundaries are assigned to domain numbers. To handle discontinuous domains comprising two ore more disjoint segments, the predicted contact map from CMAPpro is used to decide whether non-adjacent segments have a sufficient number of residue-residue contacts to be considered a single domain.

In addition to the standard 2D-RNN architectures[4,5] for the one-step prediction of entire contact maps, a variant architecture is used to predict contacts from low sequence separation (bands close to the main diagonal) to high sequence separation (bands far from the main diagonal) step by step. The predicted contact maps at lower sequence separation are used as inputs for the prediction of contact maps at higher sequence separation. The raw output of CMAPpro is a matrix of contact probabilities for all residue pairs. Several different methods for selecting contact predictions from the matrix of contact probabilities were developed and tested. Two basic methods are used in CASP 6. The first method uses a fixed threshold determined by maximizing the F-measure (harmonic mean of Precision and Recall) on a test set. The second method uses a variable, band-dependent, threshold determined by estimating the total number of contacts in a band from the sum of all the predicted contact probabilities in that band.

Our approach to tertiary structure prediction (3Dpro) combines the predicted structural features[2-5], a fragment library[11], and energy terms derived from PDB statistics. The structural features used are secondary structure, relative solvent accessibility, and a residue level contact map at a distance cut-off of 12 Å. These features are used in the energy function. A database of 9-residue fragments is constructed from the structures in the PDB. Fragments are selected from the fragment database based both on sequence similarity and similarity of the predicted secondary structure to the secondary structure of the fragment[11]. Two terms in the energy function are based directly on statistics from the PDB,

one for residue environments[11,12] and another for bond angles. To encourage the agglomeration of beta-strands into sheets we use a simple, single vector, representation of each entire strand and penalize unpaired strand vectors. We include a contact-map energy term[13], as well as a term to encourage, but not force, the secondary structure of the models to match the predicted secondary structure.

The conformational space is searched using a variant of simulated annealing, where the moves we use to modify our models are crankshaft moves[13] on one or more residues and several forms of fragment replacement[11,12]. These moves are applied to sequence locations in the model that are selected randomly. During each search, the model with the lowest energy is kept and all the other models are discarded. One thousand different models are produced using a different random seed for each search. We retain the five models with the lowest energy scores across all runs. Since our models are described in terms of the carbon alpha trace, we first add the other backbone atoms to the models, and finally use SCWRL[14] to position the side chains.


1. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. & Bourne,P.E. (2000). The Protein Data Bank. *Nucleic Acids Research* **28,** 235-242
2. Pollastri,G., Przybylski,D., Rost,B. & Baldi,P.F. (2002) Improving the Prediction of Protein Secondary Structure in Three and Eight Classes Using Recurrent Neural Networks and Profiles. *Proteins.* 47, 228-235.
3. Pollastri,G., Baldi,P.F., Fariselli,P. & Casadio,R. (2002). Prediction of coordination number and relative solvent accessibility in proteins. *Proteins* **47**, 142-153.
4. Pollastri,G. & Baldi,P.F. (2002). Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners. Proceedings of the 2002 Conference on Intelligent Systems for Molecular Biology, ISMB 02. *Bioinformatics* **18**, Supplement 1, S62-S70.
5. Baldi,P.F. & Pollastri,G. (2003). The principled design of large-scale recursive neural network architectures--DAG-RNNs and the protein structure prediction problem. *Journal of Machine Learning Research*. **4**, 575-602.
6. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.*Nucleic Acids Res.* **25**, 3389-3402.
7. Jones,D.T. (1999). Protein secondary structure prediction based on position-specific scoring matrices.*J. Mol. Biol.* 292, 195-202.
8. Przybylski,D. & Rost,B. (2002). Alignments grow, secondary structure prediction improves. *Proteins* **46**, 197-205.
9. Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., and Thornton, J.M. (1997). CATH- A Hierarchic Classification of Protein Domain Structures. *Structure* **5**, 1093-1108.
10. Pearl,F.N.G., Lee,D., Bray,J.E., Sillitoe,I. Todd,A.E.,Harrison,A.P., Thornton,J.M. and Orengo,C.A. (2000). Assigning genomic sequences to CATH. *Nucleic Acids Research* **28**, 277-282.
11. Simons,K.T., Kooperberg,C., Huang,E. & Baker,D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol.* **268**, 209-225.
12. Simons,K.T., Ruczinski,I., Kooperberg,C., Fox,B.A., Bystroff,C. & Baker,D. (1999). Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins* **34**, 82-95.
13. Vendruscolo,M., Kussell,E. & Domany,E. (1997). Recovery of protein structure from contact maps. *Fold Des*. **2**, 295-306.
14. Canutescu,A.A., Shelenkov,A.A. & Dunbrack,R.L.,Jr. (2003). A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci.* **12**, 2001-2014.


## **Bilab** - 306 models for 61 3D / 64 DR targets

### Tertiary structure prediction of proteins using assembly of flexible-length fragments and order/disorder prediction using local amino acid sequence and global alignments


S. Nakamura[1], T. Ishida[1], K. Shirakura[1], S. Mori[1], T. Terada and K. Shimizu[1]

[1] - *Department of Biotechnology, the University of Tokyo*

shugo@bi.a.u-tokyo.ac.jp


We have participated in tertiary structure prediction and order-disorder regions prediction categories in CASP6.

Disordered regions were predicted by our disorder prediction tool named “disABLE”. Our disorder prediction consisted of two steps, the prediction from local amino acid sequence and the prediction from global alignments. First, the prediction was performed by using Support Vector Machine (SVM) with position specific score matrices (PSSM) generated by PSI-BLAST, as input.

The SVM was trained with non-redundant training set generated using PISCES server[1], whose resolution cutoff was 1.6 angstrom and percentage identity cutoff was 25%, including 493 chains, 105208 residues. The predictions were performed with each three different window sizes (9, 15, and 33). The weighted average of the decision values of these predictions was calculated, and disordered regions and their reliabilities were determined by these values. Second, we searched structural templates for the target sequence by using PSI-BLAST and FFAS03 server against Protein Data Bank. If the templates included missing residues in the aligned regions, the target residues aligned to the template missing residues were judged disordered and the decision values of the residues were modified. Finally, the decision values of the prediction were denoised by low-pass filter and modified using some simple rules.

Tertiary structure prediction models for NF targets were produced by de novo protein structure modeling tool named "ABLE"[2] developed in our laboratory. For CM and FR targets, we used MODELLER to build up prediction models based on the alignments of the target and the templates obtained from fold recognition server such as 3D-PSSM, and if the target had the region without alignment we modeled the tertiary structure of such regions using by ABLE.

Modeling with ABLE was based on the general fragment assembly method and we used probability maps for mainchain torsion angles (phi-psi) at each position of the target sequence, and flexible-length fragments obtained using the match of secondary structures in addition to fixed-length (usually nine-residue) fragments. To obtain flexible-length fragments, we first search a *N*-residue fragment with similarity scores larger than a threshold. Next, we extended this fragment to *N*+1 residues and re-calculated similarity score. This process was continued until the score became lower than the threshold. The similarity score was defined by sequence identity and the match of the secondary structure. For searching flexible-length fragments, we increased the weight of the secondary structure matching to obtain longer fragments. Secondary structure prediction was performed by using PSIPRED. Typically, we could obtain fragments with more than 20 residues including multiple secondary structure elements for each NF targets. The probability maps of mainchain phi-psi torsion angles were obtained from phi-psi values of amino acids at the center of all nine-residue fragments with similarity scores larger than a threshold. For this procedure, the effects of the fragments with higher similarity scores were enhanced. Smoothing with Gaussian was applied to these maps.

After building fragment libraries and probability maps for each amino acid, 1,000-100,000 tertiary structure models of the target were produced to minimize potential energy by simulated annealing using these maps and fragments. For each simulated annealing step, the structure transition type and position were selected at random. The structural clustering was applied to produced structures and up to five structures which were the nearest from the centers of large clusters were selected. If the cluster with enough quality was not obtained, the final prediction models were selected using by some evaluation programs such as ProSa and VERIFY3D. Finally, sidechain modeling was performed for these structures by using SCWRL version 3.0.


1. Wang,G. and Dunbrack,R.L. Jr. (2003) PISCES: a protein sequence culling server. *Bioinformatics* **19**, 1589-1591.
2. Ishida,T. et al. (2003) Development of an ab initio protein structure prediction system ABLE. *Genome Inform Ser Workshop Genome Inform. 2003*; **14**, 228-237.


## **BioDec** - 67 models for 61 3D targets

### Blind-testing the entropy-filtered profile-profile alignment for fold recognition


I. Rossi[1,2], A. Zauli[1], E. Capriotti[2], P. Fariselli[2], and R. Casadio[2]

[1] - *BioDec srl, Bologna, Italy,*

[2] - *Dept. of Biology/CIRB, University of Bologna, Italy*

ivan@biodec.com


Here at CASP6 we blind-test the performance of the Entropy-filtered[1] Profile-Profile alignment method for fold recognition. This abstract summarizes the protocol used to generate the submissions for the CASP6 experiment.

Assuming that A and B are two strings of symbols, $P_A$ and $P_B$ are the rectangular matrices representing the position-specific frequency of the alphabet symbols composing the strings (superscript T indicates a matrix transpose operation), S is a (symmetric) substitution matrix, it can be derived that the matrix D, defined as $D= P^T_A S P_B$ represents the “dot” matrix for the profile comparison of the two strings. This can be efficiently computed by means of standard linear algebra routines.

For each target/template comparison, we compute the dot matrix D using the composition profiles generated by multiple alignment of the sequences reported from a five-iteration PSI-BLAST[2] search on the Non-Redundant database,

using an inclusion threshold of $E=10^{-3}$. The scoring matrix S used S is the BLOSUM62[3] substitution matrix.

Our template set comprises the structures included in the ASTRAL SCOP[5] database, release 1.65, whose sequence homology is less than 95%.

The dot matrix D is then searched for the top scoring alignment using the local Smith-Waterman dynamic programming algorithm[4]. Next, the alignments generated are subject to Shannon-entropy filtering, as described in ref.[1], using a Shannon entropy threshold of 0.5, and the remaining ones are ranked according to their Z-score. An alignment is taken into account only when its Z-score is larger than 4.

Finally, the best-ranking non-overlapping alignments are used to generate a composite CASP6 TS submission. However, if the target sequence coverage is less than 30%, the template is flagged as a putative "new fold" and a "PARENT NONE" submission is generated.


1. Capriotti,E., Fariselli,P., Rossi,I., Casadio,R. (2004) A Shannon Entropy-based filter detects high-quality profile-profile alignments in searches for remote homologues. *Proteins* **54,** 351-360.
2. Altschul,S.F. et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. **25** (17), 3389-3402
3. Henikoff,S. et al. (1998). Superior performance in protein homology detection with the BLOCKS database server. *Nucleic Acids Res.* **26**, 309-312.
4. Smith,T.S. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 147
5. Chandonia,J.M., Hon,G., Walker,N.S., Lo Conte,L., Koehl,P., Levitt,M., Brenner,S.E. (2004) The ASTRAL compendium in 2004. *Nucleic Acids Res*. **32**, D189-D192


## **BioInfo_Kuba** - 70 models for 49 3D / 21 FN targets

### Multimethod protein structure prediction


J. Pas
*BioInfoBank Institute*
kuba@bioinfo.pl


To determine whether the structure of a target protein can be predicted using homology modeling PSI-BLAST[1] search was carried out against the sequences of proteins in the non-redundant protein sequence. PSI-BLAST iterations were performed using manual inclusion/exclusion procedure.

After that multiple sequence alignment was built using clustalw[2] program using selected proteins from PSI-BLAST profile. All alignments were manually inspected.

Selection of template was confirmed using structure prediction METASERVER[3]. METASERVER was also used to choose template when no significant hits were found using PSI-BLAST searches.

In addition other available information was used in an attempt to link the target with a protein with known structure. It was mainly literature search, known metabolic pathways, gene expression data, position on the chromosome, distribution of folds in the organism and secondary structure prediction.

Selected target–template structural alignments were visually inspected in SWISS PDB Viewer and if necessary modified. Molecular 3D models were then built 3D using both SWISS-MODEL[4] and MODELLER[5] programs. Initial models were subjected to detailed evaluation, mainly by addition visual inspection of structural consistency and using Verify 3D program[6]. The same evaluation procedure was performed for final models.

More than one template protein was used if possible after superimposition of their molecular structures using 3d-hit program[7]. During the modeling procedure superimposition of initial models were used to find best possible backbone conformation

The overall quality of each modeled structure was evaluated in detail with the Verify 3D program.

1. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
2. Thompson,J.D., Higgins,D.G., Gibson, TJ. (1994). CLUSTAL W: improving the sensivity of progressive multiple sequence alignment through sequence weighting. *Nucleic Acids Res.* **22**, 4673-4680.
3. Bujnicki,J.M., Elofsson,A., Fischer,D., Rychlewski,L. (2001). Structure prediction meta server. *Bioinformatics* **17**, 750-751.
4. Guex,N., Peitsch,M.C. (1997). SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* **18**, 2714-2723.
5. Sali,A., Blundell,T.L. (1993). Comparative protein modeling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779-815.
6. Luthy,R., Bowie,J.U., Eisenberg,D. (1999). Assessment of protein models with three-dimensional profiles. *Nature* **356**, 83-85.
7. Plewczynski,D., Pas,J., von Grotthuss,M., Rychlewski,L. (2002). 3D-Hit: fast structural comparison of proteins. *Appl Bioinformatics* **1**, 223-225.

## **Biopred** - 64 models for 62 FN targets

### Annotate CASP6 targets using sequence and structural information

G. Cheng[1], D. Baker[1], and Ram Samudrala[2]
*1 -Department of Biochemistry, University of Washington,*
*2 -Department of Microbiology, University of Washington*
ram@compbio.washington.edu, {gcheng,dabaker}@u.washington.edu

We functionally annotate CASP targets using a combination of automated annotations derived from the Bioverse database and webserver (http://bioverse.compbio.washington.edu) as well as manual ones based on all the information we could collect including sequence, structure and literature.

The Bioverse automated framework for annotation is described elsewhere[1]. For the manual approach, we first perform a psi-blast search[2]. If this yields well-annotated sequences, the GO function annotations of the homologues are simply inherited. The GO process and GO component annotation are based on literature related to the sequence. If psi-blast doesn't give enough information, then the Sanger center pfam annotation web server[3] is used to annotate the target sequence. If pfam hit is a DUF (Domain of Unknown Function), the 3D Jury template from bioinfo.pl[6] is used to judge the function of the sequence. Because the 3D Jury score represents the structural similarity between the target and template best, we still check whether each template sequence profile has functional motif residues that are aligned well with the target sequence profile to ensure the function is not changed during evolution. If both template and target have similar conservation pattern and the 3D Jury score is high, we assign the function of target based on the best aligned 3D Jury template. For ab initio targets, we have developed a motif based search algorithm to search the decoy ensemble based on the motif to identify the potential function of the target[4].

The function site, binding site predictions are also based on the characteristics of the target sequence. For psi-blast level targets, we do homology based binding site mapping. The algorithm will superimpose[5] the ligand from the template PDB on the homology model[7-8] and map the functional residue based on parent. For FR level targets, the same algorithm is used to map the function site, but the FR model and the FR template are chosen based on function site conservation. During the function site mapping, we can often discover alignment problem at FR level targets. This process also helps us build better FR level models. For *ab initio* targets, the function site is mainly predicted by sequence conservation, since there are only very few such targets. Those conserved polar residues clusters are identified manually. For some targets, metal binding sites could be visually identified from the sequence alignment based on residue type and sequence conservation.

1. McDermott,J., Samudrala R. (2003). BIOVERSE: Functional, structural, and contextual annotation of proteins and proteomes. *Nucleic Acids Res.* **31**, 3736-3737.
2. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W., Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389-3402, 1997.
3. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M., Sonnhammer,E.L. (2002). The Pfam protein families database. *Nucleic Acids Res.* **30**, 276-280.
4. Cheng,G., Samudrala,R., Baker,D. (2004). Unpublished results.
5. Ortiz,A.R., Strauss,C.E., Olmea,O. (2002). MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.* **11**, 2606-2021.
6. Ginalski,K., Elofsson,A., Fischer,D., Rychlewski,L. (2003). 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics* **22**, 1015-1018.

7. Hung,L.H., Samudrala,R. (2003) PROTINFO: Secondary and tertiary protein structure prediction. *Nucleic Acids Research* **31**, 3296-3299.
8. Kim,D.E., Chivian,D., Baker,D.. (2004) Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res.* **32,** W526-31.

## **Biovertis** - 52 models for 49 3D targets

### Protein structure prediction pipeline for industrial research

Walter A. Koppensteiner
*Biovertis – Information driven drug design AG, Campus Vienna Biocenter 6, A-1030, Vienna, Austria*
Walter.Koppensteiner@biovertis.com

Biovertis deploys, among other bioinformatics techniques, protein structure prediction for the identification and validation of novel anti-microbial drug targets. Structure prediction serves mainly two purposes: (1) inferring the function of uncharacterized proteins and (2) building structure models to accelerate subsequent NMR structure determination. All atom predictions are modeled when required for certain applications like docking.

We have established a pool of prediction techniques to reduce the dependency on the strengths and weaknesses of a particular method. Thus our pipeline integrates both sequence profile and threading methods complemented by a set pre-processing and post-processing techniques. Subsequently, our prediction pipeline will be presented, where the actual structure prediction is separated from pre- and post-prediction steps.

*Pre-prediction steps*
For each target protein, we predict the secondary structure[1], transmembrane helices[2], signal peptides[3], low complexity regions[4] and coiled coils[5]. Additionally, we use InterProScan[6] to identify InterPro[7] signatures in the target sequence. The predicted features support the subsequent steps but also allow making a statement about function and putative domain borders. Moreover, we have observed that hydrophobic regions may confuse threading algorithms and we remove such stretches from the sequence.

*Structure prediction*
The initial method deployed is an iterative sequence search using PSI-Blast[8]. The first PSI-Blast run searches NCBI's non-redundant (nr) sequence database. The second run reloads the checkpoint file and searches a sequence database of known structures, either a domain database derived from SCOP[9] or sequences from the proteins in PDB[10]. Both sequence databases are clustered with a threshold of 95% sequence identity to remove redundancy.

If PSI-Blast does not allow us to make a satisfactory prediction we switch to the structure based methods FUGUE[12] and ProFit[11] which are applied simultaneously. The latter uses the same fold libraries as PSI-Blast, FUGUE uses the HOMSTRAD[13] database.

All structure prediction results undergo visual inspection of both alignments and 3D models. Here we also incorporate the features predicted prior to structure prediction. We regard this step as essential and avoid automated assignments because we can eliminate false positive hits, can identify domain boundaries and can recognize gross alignment errors.

In many cases, structure prediction is not completed in one cycle. Instead, multi-domain proteins may undergo several repetitions where each domain is predicted independently using the sequence of the putative domain as input. It is thus not unusual, that different domains of a protein have been predicted with different methods.

*Post-prediction steps*
Once a structural template has been found for a domain by one of the methods described above, we have some post-processing methods to our disposal. If the decision for the best template is ambiguous or the alignment quality appears unsatisfactory, we deploy Prosa[14] to base the optimization on the z-scores and the energy profiles of the 3D models. Variations of the alignment can be generated through the adjustment of ProFit parameters or by hand. If an all-atom model is desired and the alignment quality allows the construction of a sufficiently accurate model, we deploy third party software to construct such a model. For CASP6 we have restricted our efforts to the prediction of the best template for cases where sequence similarity was marginal.

*Environment and packages*
For good reasons, we have not implemented a "one-script-does-everything" approach. Instead, we have built an environment where we can use the methods in a flexible manner and which allows human intervention. For structure prediction we have licensed the packages ProHit from ProCeryon Biosciences GmbH, which integrates PSI-Blast and ProFit, and Tripos' Sybyl which contains FUGUE and HOMSTRAD. For routine applications, parameters are set to their default values. If this does not give a satisfactory result, we change parameters to increase sensitivity.

*Summary of results*

For CASP6 we have submitted predictions for 60 targets, of which 34 have been predicted with PSI-Blast. The remaining predictions were based on ProFit and FUGUE results, often in a consensus manner. We aimed to make an unambiguous prediction for every target. For three targets, however, we made use of the option to submit more than one model per target. For another three targets, we submitted a model containing two domains. All other models were single domain predictions.


1. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* ***292***, 195-202.
2. Krogh,A., Larsson,B., von Heijne,G. & Sonnhammer,E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567-80.
3. Nielsen,H., Engelbrecht,J., Brunak,S. & von Heijne,G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng*. **10**, 1-6.
4. Wootton,J.C. & Federhen,S. (1993) Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem*. **17**, 149-163.
5. Lupas,A., van Dyke,M. & Stock,J. (1991) Predicting coiled coils from protein sequence. *Science* **252**, 1162-1164.
6. Zdobnov,E.M. & Apweiler,R. (2001) InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847-8.
7. Mulder,N.J. et. al. (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res*. **31**, 315-318.
8. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
9. Murzin,A.G., Brenner,S.E., Hubbard,T. & Chothia,C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536-540.
10. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. & Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.* **28**, 235-242.
11. Sippl,M.J. & Weitckus,S. (1992) Detection of native like models for amino acid sequences of unknown three dimensional structure in a data base of known protein conformations. *Proteins* **13**, 258-27.
12. Shi,J., Blundell,T., & Mizuguchi,K. (2001) FUGUE: Sequence-Structure Homology Recognition Using Environment-Specific Substitution Tables and Structure-Dependent Gap Penalties. *J. Mol. Biol.* **310**, 243-257.
13. Mizuguchi,K, Deane,C.M., Blundell,T.L., & Overington,J.P. (1998) HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.* **7**, 2469-2471.
14. Sippl,M.J. (1993) Recognition of Errors in Three-Dimensional Structures of Proteins. *Proteins* **17**, 355-362.


## **Bishop** - 150 models for 30 3D targets

# A simple approach for ab-initio protein structure prediction


Shing-Chung Ngan, Ling-Hong Hung and Ram Samudrala

*Dept. of Microbiology, University of Washington, Seattle, WA, USA*

ngan@compbio.washington.edu


Introduction

Our approach for ab-initio structure prediction for CASP-6 consists of three steps: (i) exploring the conformational space based on simulated annealing and appropriate energy functions, in order to generate a set of decoys (10000-20000 conformations) for a protein sequence of interest, (ii) filtering the decoys using various energy functions, with the goal of enriching the overall quality of the remaining decoys (300-600 conformations), and (iii) visually inspecting the distribution of those remaining decoys through multi-dimensional scaling, in order to look for clusters, and to pick five final conformations from those cluster centers. The overall framework of the approach is simple. Novelties lie in the choice, construction, and combination of the energy functions, and in the formation of the hierarchical filters.

Step 1. Conformational Space Exploration

Residues predicted with high confidence to be part of a helix or sheet are first set to the idealized helix and sheet phi-psi values. The rest of the residues in the protein chain are set to extended conformation. Then, a standard Monte Carlo scheme with simulated annealing is used to modify the conformation of the "non-high-confidence" residues, by perturbing consecutive triplets of residues at random positions. The perturbation of the triplet conformation is based on the standard fragment replacement scheme[1]. The overall energy function used in simulated annealing is a combination of six energy functions: (1) hydrophobic compactness, (2) bad-contacts penalty, (3) an all-atom distance-dependent pair potential[2], (4) a residue-based distance-dependent triplet potential, (5) a phi-psi potential, and (6) a potential based on the radial distance of a residue from the center of the conformation. (The relative weights for these energy functions were predetermined by applying an iterative training procedure on a set of test

proteins.) Around 1000 to 2000 seeds are used to generate 10000-20000 decoys.

Step 2. Filtering

Our goal is to filter the 10000-20000 decoys down to a smaller set of decoys with better quality. To achieve this, various linear combinations of a set of energy functions are applied on the decoy set in a hierarchical manner. (The weights used in the linear combinations were derived based on performing logistic regression hierarchically on various subsets of the test proteins. A total of 13 hierarchical filters were constructed.) This set of energy functions includes the six mentioned in the previous section, plus three physical functions (electrostatics, Van der Waals, and solvation), and several probabilistic functions (virtual torsion angle, solvation state, a residue pair potential taking into account the degree of conservation of the residues, a potential based on the probability of a residue being within a prescribed cutoff distance from other hydrophobic, hydrophillic and neutral residues, etc.). 300-600 decoys are retained at the end of this step.

Step 3. Selection

Multi-dimensional scaling is used to produce a reduced-dimensional plot, which enables us to visually observe the distribution of the remaining 300-600 decoys. In this plot, a particular decoy can be represented by more than one point if it was picked up by more than one hierarchical filter in Step 2. (The goal is to preserve possible consensus information available among the filters.) We look for clusters of decoys and select five final conformations from the centers of these clusters. This final step of visually picking five conformations, a non-automated process, can be replaced by an automated scheme such as statistical clustering in the future.


1. Simons,K.T., Kooperberg,C., Huang,E., Baker,D. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J. Mol. Biol.* **268**, 209-225.
2. Samudrala,R., Moult,J. (1998) An all-atom distance-dependent conditional probability discriminatory functions for protein structure prediction. *J. Mol. Biol*. **275**, 893-914.


**BMERC** (serv) - 133 models for 44 3D targets

## PSDM: a tool for protein structural domain prediction using bayesian fold recognition


Praveen F. Cherukuri[1], Gregory D. McAllister[2], Temple F. Smith[3] and Jadwiga R. Bienkowska [2,3]

[1] *Bioinformatics Program, Boston University, MA*

[2] *Serono Reproductive Biology Institute, One Technology Place, Rockland, MA*

[3] *BMERC, Boston University, MA*

Jadwiga.Bienkowska@serono.com


Our structure prediction method employs Bayesian Fold Recognition (BFR) and sequence-profile alignment software PIMAII. See Figure 1 for an overview of how the method works. The server requires: (a) primary protein sequence in FASTA format and (b) an e-mail address where the results are to be sent. The current implementation of the BFR assumes that the query sequence represents one structural domain. It is recommended that prior to submission to the structure prediction server, the sequence of a multi-domain protein is analysed by available tools such as CDART[1] and the sequence of putative single domains are submitted to PSDM separately. This structure prediction service is available as a publicly accessible web-based tool at http://bmerc-www.bu.edu/cgi-bin/pcheruku/ServerScripts/PSDM.cgi.

The first step in our approach uses BFR to select a set of fold models most compatible with a query sequence. Currently BFR uses a library of over 20,000 automatically built DSMs[2-4]. This DSM library was constructed from all protein domains classified in the SCOP database[5], release 1.61, which have less than 95% sequence identity[6] and an additional set of protein structures that are not yet classified in SCOP and have less than 40 % identity among themselves.

The BFR uses a filtering algorithm[7,8] to calculate the probability of the sequence given a DSM model of a structural domain. Typically several domain models represent each distinct fold. Thus fold model is equivalent to the set of models of domains that are classified under the same fold. The BFR assigns a posterior probability to each fold. For each fold one structural domain is selected as the best fold representative for the query sequence. Different query sequences may have different SCOP domains selected as the best fold representative even though the fold prediction is the same.

The query sequence is first threaded through the DSM library and the probability of observing a sequence given the model, P(seq | model), is

calculated by the Filtering algorithm developed by White[7] . Each fold is represented by several DSMs and the prior probability of observing a sequence given a fold model is:

$$P(seq \mid fold_i) = \max\{P(seq \mid model_k), model_k \in fold_i)\} \qquad (1)$$

The posterior model probability given the query sequence calculated according to the Bayesian formula[3] is :

$$P(fold_i \mid seq) = \frac{P(seq \mid fold_i) \cdot P(fold_i)}{\sum_{j=1}^{k} P(seq \mid fold_j) \cdot P(fold_j)} \qquad (2)$$

where $P(fold_i)$ is the prior probability assigned uniformly to $fold_i$ over $k$ folds. Alternative methods for priors assignment have been investigated and it has been established that they lead to insignificant improvements in the fold recognition performance[9] over the method implemented here. Thus for a query sequence the BFR selects the best representative SCOP domain for each fold and assigns the posterior probability associated with each fold.

We recommend a use of a binary decision rule and the top ranking fold is considered as an acceptable prediction if its posterior probability is greater than 0.5. Nevertheless, up to top 5 models with posterior probability greater than 0.01 are selected for further analysis by PIMAII.

The second step involves PIMAII and assesses the similarity of the query sequence with the primary sequence of the selected SCOP domains. PIMAII aligns profile- defining sequences for the selected SCOP functional domains and the query sequence. Each profile is defined by a set of homologous domain sequences. PIMAII is an iterative local dynamic programming alignment algorithm described previously in[10,11] which begins with the two most similar sequences and identifies a locally optimal alignment using the scoring scheme described in[11]. The resulting profile of these two most similar sequences is then aligned to the next most similar sequence; the procedure is continued until all the sequences have been aligned and a single alignment matrix is obtained. A sequence may be skipped in any step of the alignment process if the information content of the generated profile drops below a predetermined value. We have tested different values of parameters used by PIMAII to optimize this approach for the alignment of sequences with low similarity and have selected a value of 5.0 for the information-content-cutoff [11] .

In order to align a query sequence to a SCOP domain sequence we use a set of SCOP domain homologs. Homologs of a domain are selected through an automated procedure described previously in[4]. Homologous sequences are pooled together with the query sequence and PIMAII is used to generate a common profile and alignment. If for any of the selected top 5 folds the alignment can be generated, the server reports results. If the profile alignment fails for all 5 top folds a pairwise sequence alignment is generated between the query and the sequence of the top fold domain.


1. Geer,L.Y. et al. (2002). CDART: protein homology by domain architecture. *Genome Res.* 1619-23.
2. Bienkowska,J.R., Rogers,R., Yu,L. and Smith,T.F. (2000). Protein fold recognition by total alignment probability. *Proteins* 451-62.
3. Bienkowska,J.R., Rogers,R., He,H. and Yu,L. (2002). Bayesian approach to protein fold recognition. *Protein structure prediction. Bioinformatics approach*, I. Tsigielny, Editor. IUL.
4. Bienkowska,J.R., He,H. and Smith,T.F. Automatic Pattern Embedding In Protein Structure Models. *IEEE*. 2001. 21-25.
5. Lo Conte,L., et al. (2000). SCOP: a structural classification of proteins database. *Nucleic Acids Res.* 257-9.
6. Brenner,S.E., Koehl, and Levitt,M. (2000) The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res*. 254-6.
7. White,J.V. (1988). Modeling and filtering for discretely valued time series. Bayesian analysis of time series and dynamic models, S. J.C, Editor., Marcel Dekker: New York. 255-283.
8. Rabiner,L.R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc IEEE*. 257-286.
9. He,H., McAllister,G. and Smith,T.F. (2002). Triage protein fold prediction. *Proteins*. 654-63.
10. Smith,R.F. and Smith,T.F. (1990). Automatic generation of primary sequence patterns from sets of related protein sequences. *Proc Natl Acad Sci U S A*. 118-22.
11. Das,S. and Smith,T.F. (2000). Identifying nature's protein lego set. Advances in Protein Chemistry, .P.S. Kim, Editor. AP Press: New York. . 159-183.

**boniaki_pred** - 272 models for 56 3D targets

## Protein structure prediction using intermediate resolution off-lattice protein model (Refiner) and public accessible threading and comparative modeling tools

M.J. Boniecki[1,2], A. Koliński[2], J.M. Bujnicki[1]
*[1] – International Institute of Molecular and Cell Biology, Laboratory of Bioinformatics and Protein Engineering, [2] – Warsaw University, Faculty of Chemistry, Laboratory of Theory of Biopolymers*
mboni@genesilico.pl

At first each protein sequence was submitted to the Genesilico meta-server[1]. The meta-server sends sequence to several servers which generate sequence alignments. The Genesilico meta-server receives alignments, names of templates and scores from servers. The meta-server additionally scores alignments using Pcons[2] and generates a set of rudimentary models numerated according to the target.

Further path of prediction depended on scores, consistence of alignments and consistence of rudimentary models. Models were compared and consensus distance restraints were calculated. For each pair of residues, depending on distance consensus and distance between these two residues, energy weight, and distance range without penalty, were calculated.

In cases of high scores and good consistence of restraints and alignments, comparative modeling approach were performed. Initial model was generated using standard comparative modeling tools, in the most cases Modeller[3], in the remaining cases Swiss-Model[4]. In cases of especially good scores, assigned to only one template, high sequence similarity, these fragments, sometimes almost entire protein, were fixed. Remaining fragments were refined or reconstructed using Refiner. The reconstruction usually involved restraints obtained from crude models. In cases, in which there was several good hits, all of them were used to calculated restraints and entire molecule was refined by Refiner[5].

In cases of not very good but reasonable scores and models that seemed to be consistent, consensus distance restraints were calculated. When restraints showed some consistence they were used in refining (refolding) simulation, performed by Refiner. Starting structure was obtained using comparative modeling tools, from best scored (or continuous) alignment. In some cases simulation started from extended structure and it was fold using restraints.

In cases of bad scores and crude models that don't show sufficient consistence, an ab initio folding procedure was applied. In such cases Refiner was started from extended structure without any restrains. When sequence was too long to be treated by Refiner, protein seemed to be multidomain an attempt to get some additional information was made. Usually Baker-Robetta[6] or public accessible CAFASP models were used to achieve some restraints.

Refiner is an off-lattice intermediate resolution protein model. It represents protein as a chain of Cα atoms connected to each other using virtual bonds of constant distance 3.8Å. Side chains are represented by one or two united atoms, depending of size of the side group. Refiner is a program based on energy minimization. It employs complex statistical forcefield calculated from database of native structures. Refiner's conformational searching scheme is based on Monte Carlo methods. It employs asymmetrical Metropolis scheme embedded in Replica Exchange Monte Carlo scheme.

Simulations were calculated on computer clusters in Interdisciplinary Center for Mathematical and Computational Modeling (Warsaw University) and Laboratory of Bioinformatics and Protein Engineering (International Institute of Molecular and Cell Biology – in Warsaw).

1. Kurowski,M.A., Bujnicki,J.M. (2003). GeneSilico protein structure prediction meta-server. *Nucleic Acids Res.* **31**(13), 3305-3307.
2. Lundström,J., Rychlewski,L., Bujnicki,J., Elofsson,A. (2001) Pcons: A neural-network-based consensus predictor that improves fold recognition. *Protein Sci.* **10**(11), 2354-2362.
3. Sali,A., Blundell,T.L. (1993) Comparative modelling by satisfaction of spatial restraints. *J. Mol. Biol*. **234**, 779-815.
4. Guex,A., Peitsch,M.C. (1997) Swiss-Model and the Swiss-Pdb-Viewer an enviroment for comparative protein modeling. *Electrophoresis* **18**, 2714-2723.
5. Boniecki M., Rotkiewicz P., Skolnick J., Kolinski A. (2003) Protein fragment reconstruction using various modeling techniques. *J. Computer Aided Molecular Design* 17, 725-737.
6. Chivian,D., Kim,D.E., Malmstrom,M., Bradley,P., Robertson,R., Murphy,P., Strauss,C.E., Bonneau,R., Rohl,C.A. Baker,D. (2003) Automate prediction of CASP-5 structures using the Robetta server. *Proteins* **53**, 524-533.

## **Brooks-Zheng** - 170 models for 38 3D targets

### Generate models for new folds by rewiring helices and strands from known protein structures

Wenjun Zheng and Bernard Brooks
*National Heart, Lung, and Blood Institute,*
*National Institutes of Health, Bethesda, MD 20892*
zhengwj@helix.nih.gov

We propose a highly efficient way of generating structural models for potentially 'new fold' targets from known protein structures as follows: based on the secondary structure prediction, the target sequence of a single domain is partitioned into segments of helices and strands; each segment is aligned onto a given template structure based on compatibility of local hydrophobicity and secondary structures; then all the aligned segments are 'rewired' by loops connecting between neighboring helices or strands.

There are several major advantages: first, by allowing a 'rewiring' procedure that does not follow the template's sequential order (plus a swapping of C-terminal and N-terminal of each segment) new folds with novel topologies can be generated easily; secondly, by directly copying coordinates of helices and strands from known structures, 'native like' features are kept in the generated models (for example: well-formed beta sheets, native-like super-secondary structural contacts etc); thirdly, by focusing on the packing of helices and strands while postponing the loop-modeling to a latter stage, we can reduce the search space significantly; fourthly, by extracting non-ideal helices/strands from the existing PDB database we may be able to model twisting and bending features of realistic secondary structures.

We conducted a preliminary study on the feasibility of generating a new fold by rewiring the helices and strands from another old fold (fold definitions based on Dali [1]). Indeed, depending on classes of secondary structural content, 40%-75% of them can be modeled with cRMSD=5Å by this method.

This method in principle generalizes the standard threading-based fold recognition algorithms by adding new flexibilities to model generation with minimal increase of computational cost. By applying a set of selected score functions and then clustering procedure to the generated models, we expect to select a close-to-native model for the packing of helices and strands. Further modeling of the loops will be pursued in the future.

1. Holm, L. & Sander C (1998). Touring protein fold space with Dali/FSSP. *Nucl. Acids. Res.* **26**, 316-319

## **BUKKA** - 82 models for 18 3D targets

### *Ab initio* potein sructure prediction

D. Katagiri, H. Ode, H. Ishikawa, T. Hattori,
Y. Syoji and T. Hoshino
*Graduate School of Pharmaceutical Science, Chiba University*
k-dai@graduate.chiba-u.jp

A theoretical method such as ab initio quantum chemical calculation and molecular dynamics (MD) simulation is widely used for the analysis on functional and structural properties of biomolecules, and the biomolecular functions can be qualitatively evaluated[1]. As for the local structures of proteins, it was suggested from quantum chemical techniques why secondary structures like helix and β-sheet were dominantly stabilized.[2, 3] However, it is difficult to predict the whole protein structures and further to quantitatively evaluate the binding capacity through the theoretical method. In present, the theoretical method has two problems. First, a large calculation time is needed for quantum calculation even by the latest computational equipment. Second, the currently available force field is incomplete to express the protein structures consisting of amino acids in spite of considerable accumulation of force fields, those have been derived from the post ab initio calculations or experimental data for several small molecules. It seems that a new force field which is based on ab initio calculation for all kind of amino acid residues is needed to MD simulation. This new force field would lead us to catch the global minimum structure of proteins, reasonably.

Our protein structure predictions were carried out for 24 targets that have 130 or less residues and including two canceled targets. All structure predictions were started from a straight form of the polypeptide chain as the initial configuration. Then, a temperature ramp was used to suddenly raise the temperature of the whole system up to 500K for 80 ps. After this heating procedure, cooling simulation was performed. A temperature ramp was used to gradually decrease the temperature of the whole system down to 288 K for 7 ns. All molecular dynamics simulations were performed with a 1.0 fs time step, a no cut off for Lennard-Jones interactions, and the use of SHAKE[4] for restricting

motion of all covalent bonds involving hydrogen, using modified version of the AMBER 7[5] suite of programs.

As for a force field, originally developed force field was employed, where each of the 20 amino acids has the respective parameters set. The parameters for the 20 amino acids were generated by the force field parameterizing technique developed by us, in which quantum chemical calculation is essentially required. Accordingly, the structures of amino acids were optimized by Gaussian 98[6] program using density functional method[7] ( B3LYP ) with 6-31G** basis set, before generating the respective force field parameters.

Solvent effects were incorporated using the Generalized Born model[8], as implemented in AMBER 7.

The structures of 5 models, T0196[9], T0205[10], T0207[11], T0212[12] and T0254[13], have been registered on the Protein Data Bank[14] (PDB) in 13-Oct-2004. The backbone rmsd between PDB structures and the prediction structures were T0196 - 15.16 Å, T0205 Chain A - 11.45 Å, T0205 Chain B – 15.83, T0207-11.63 Å, T0212-19.36 Å, T0254 Chain A – 16.65 Å and T0259 Chain B – 16.66 Å, respectively. In the T0196 molecule whose local structure corresponding to the 90 residues is cleared in PDB, 28 residues of the predicted structure matched with PDB from the view point of the secondary structure. The secondary structure was defined by classifying the residues into helix, β-sheet and other structure using dssp[15] program. In the T0205 Chain A, T0205 Chain B, T0207, T0212, T0254 Chain A and T0254 Chain B molecule, the number of 19 out of 69, 16 out of 103, 31 out of 75, 22 out of 126, 42 out of 107 and 42 out of 107 residues were compatible with PDB, respectively. Calculated potential energies of the predicted structures were stable than those of PDB structures in all case.

Our prediction accuracy was high for the helix region, however, low for the β-sheet region. β-sheet region tended to be predicted as amorphous structure, and turn region between β-sheet and β-sheet structure, as helix structure. Moreover, the prediction accuracy was low in the region including a lot of polarity amino acid side chains such as Arg, Asp, Glu and Lys. It is quite likely that these low accuracies depend on the computational condition of making the original force field in vacuo. As a matter fact, structural difference between in water and in vacuo was observed[16]. Therefore, making the original force field in water solvent will be needed to achieve a higher accuracy prediction.


1. Katagiri,D., Hata,M., Itoh,T., Neya,S. & Hoshino,T. (2003) Atomic Scale Mechanism of the GTP → GDP Hydrolysis Reaction by the Giα1 Protein. *J. Phys. Chem. B.* **107**, 3278-3283.
2. Katagiri,D., Tsuchiya,T., Tsuda,M., Hata,M. & Hoshino,T. (2002) Computational Analysis of Stability of the β-sheet Structure. *J. Phys. Chem. B.* **106**, 9151-9158.
3. Tsuchiya,T., Katagiri,D., Hata,M., Hoshino,T. & Tsuda,M. (2002) Theoretical analysis of the stability of helices. *J. Mol. Str. (Theochem).* **589-590**, 413-422.
4. Ryckaert,J.P., Ciccotti,G. & Berendsen,H.J.C. (1977) Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *J. Computat. Phys.* ***23***, 327-341.
5. Case,D.A. et al. (2002) AMBER 7, University of California, San Francisco.
6. Frisch,M.J. et al. (1998) *Gaussian98*, revision A.7; Gaussian,Inc.: Pittsburgh,PA.
7. Becke,A.D. (1993) Density-functional thermochemistry. III; The role of exact exchange. *J. Chem. Phys.* **98**, 5648.
8. Tsui,V. & Case,D.A. (2001) Theory and applications of the generalized Born solvation model in macromolecular simulations. *Biopolymers. (Nucl. Acid. Sci.)* **56**, 275-291.
9. Chang,J.C., Zhao,M., Zhou,W., Liu,Z.J., Tempel,W., Arendall III,W.B., Chen,L., Lee,D., Habel,J.E., Rose,J.P., Richardson,J.S., Richardson,D.C. & Wang,B.C. Hypothetical Protein from Pyrococcus Furiosus Pfu-880080-001 To be Published.
10. Wesenberg,G.E., Smith D.W., Phillips Jr,G.N., Bitto,E., Bingman,C.A. & Allard,S.T.M. X-Ray Structure of Gene Product from Arabidopsis Thaliana at2G34160 To be Published.
11. Kuzin,A.P., Vorobiev,S.M., Lee,I., Edstrom,W., Acton,T.B., Ho,C.K., Cooper,B., Ma,L.-C., Xiao,R., Montelione,G., Tong,L. & Hunt,J.F. X-Ray Structure of Northeast Structural Genomics Target Protein Xcr50 from X. Campestris To be Published.
12. Kuzin,A.P., Vorobiev,S.M., Edstrom,W., Acton,T.B., Shastry,R., Ma,L.-C., Cooper,B., Xiao,R., Montelione,G., Tong,L. & Hunt,J.F. X-Ray Structure of Northeast Structural Genomics Consortium Target Sor45 To be Published.
13. Binkowski,T.A., Wu,R., Moy,S.F. & Joachimiak,A. Hypothetical Protein from Vibrio Cholerae O1 Biovar Eltor Str. N16961 To be Published.
14. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N., Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res*. **28**, 235-242.

15. Kabsch,W. & Sander,C. (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical feature. *Biopolymers*. **22**, 2577-2637.
16. Wang,Z.X. & Duan,Y. (2004) Solvation Effects on Alanine Dipeptide: A MP2/cc-pVTZ//MP2/6-31G** Study of (Φ,Ψ) Energy Maps and Conformers in the Gas Phase, Ether, and Water. *J. Comput. Chem*. **25**, 1699-1716.

## **Bystroff** - 36 models for 30 RR targets

### Contact map prediction using HMMSTR

X. Yuan[1], Y. Hou[2], Y.-M. Huang[1], Y. Shao[1] and C. Bystroff[1]
*[1] - Dept of Biology, Rensselaer Polytechnic Institute, Troy, NY 12180,*
*[2] - Dept of Computer Science, National University of Singapore, Kent Ridge, Singapore 119260*
bystrc@rpi.edu

Predictions were made in the residue-residue contact (RR) format for CASP6. Sequences that were determined not to have a close homolog among the known structures were predicted using HMMSTR, a hidden Markov model for local structure motifs[2], and associated algorithms.

A PSI-BLAST[1] amino acid profile for the target was used to calculate the position-specific Markov state probabilities, gamma. The gamma matrix was classified into one of 58 SCOP superfamilies using a support vector machine, SVM-HMMSTR, that was trained for fold recognition using HMMSTR state composition and local dynamic programming alignments[3]. The gamma matrix was used to calculate pairwise contact potentials, Eij, using a method described previously[4].

To predict contact maps from contact potentials, one of the following three approaches was used: (1) Alignments were made between the target Eij matrix and template Eij matrices of SVM-HMMSTR hits, using a fragment assembly approach. Or, (2) conserved features were recognized by eye in the colorized Eij image and template images or SVM-HMMSTR hits, and alignments were drawn by hand. Or, (3) no templates were used and contacts were predicted directly from the target Eij image using pathways.

Possible folding pathways were designed by assigning contacts first to local supersecondary structures with good Eij scores, then to protein-like contact map features that were physically possible given the already-defined contacts. Simple rules and drawings were used to predict which contacts were physically possible. Combinations of template-based and *ab initio* predictions were sometimes made. Each target had a different story, and the strategy evolved over the course of the CASP6 season. More on this method can be found from the following URL: http://www.bioinfo.rpi.edu/~bystrc/pub/casp6abstract.pdf.

1. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
2. Bystroff,C., Thorsson,V. & Baker,D. (2000). HMMSTR: A hidden markov model for local sequence-structure correlations in proteins. *J. Mol. Biol.* **301**, 173-190.
3. Hou,Y., Hsu,W., Lee,M.L., Bystroff, C. (2004) Remote homolog detection using local sequence-structure correlations. *Proteins* **57(**3):518-30.
4. Shao,Y & Bystroff,C. (2003). Predicting inter-residue contacts using templates and pathways. *Proteins* **53** Suppl 6, 497-502

## **CAFASP-Consensus** - 64 models for 64 3D targets

### CAFASP-CONSENSUS

H.K. Saini and D. Fischer
*Center of Excellence in Bioinformatics, University at Buffalo*
*901 Washington St., Suite 300, Buffalo, NY 14203*
hkaur@bioinformatics.buffalo.edu

The idea of CAFASP-CONSENSUS was to submit predictions that corresponded to publicly available information published upon the release of each target at CAFASP web site (http://www.cs.bgu.ac.il/~dfischer/CAFASP4/targets.html). The published information corresponded to the selection of 3djury, which produces a consensus prediction using the models reported by all the CAFASP servers. In many cases, the selection consisted of only $C^{\alpha}$, gappy models. In order to generate full-atom refined models, the 3djury selection was refined using the Nest package. The CAFASP-CONSENSUS predictions to CASP thus entail a baseline level of performance to which human CASP predictors could be compared to.

## **CaspIta** - 348 models for 64 3D / 63 DP / 63 DR / 64 FN targets

### An integrated approach to the prediction of protein structure and function

S.C.E. Tosatto[1], O. Bortolami[1], A. Cestaro[1], G. Cozza[2], M. Lexa[1], S. Toppo[3], G. Valle[1], and S. Moro[2]
*[1] – Dept. of Biology and CRIBI Biotech Centre, [2] – Dept. of Pharmaceutical Sciences, [3] – Dept. of Biological Chemistry, University of Padova*
silvio@cribi.unipd.it

We describe a method integrating the major aspects of protein structure and function prediction. The process starts with simple 1D predictions which are used to simplify the following steps leading towards full 3D prediction. In some cases results gathered further down the prediction pipeline are used to improve the initial decisions.

The first step for each target consisted in predicting stretches of protein disorder. This was done using an experimental SVM (support vector machine) method trained to discriminate disordered regions according to their sequence composition. A sliding window of 11 residues is used to calculate the reative abundance for each of the 20 amino acids. This SVM was trained on two sets of ordered, resp. disordered, sequences. Since the SVM has a tendency to overpredict disordered regions, the predicted secondary structure and, in the case of homology modeling targets, presence of a structural template were used to remove false positives.

The domain structure of each target was then predicted based on a combination of several methods. First, the CDD database[1] is scanned to determine obvious domain boundaries. Longer sequence fragments predicted to be disordered are also excluded from further analysis. PEPTIMEX, an amino acid extension of the PRIMEX method[2] for sequence pattern matching, is then used as an experimental ab initio method for domain identification. The program uses correlated sequence patterns at a given distance to establish the likelihood of two fragments belonging to the same domain. In some fold recognition cases the domain structure was also inferred from the presence of particular folds covering part (or all) of the target sequence.

The function of each target domain was assessed in terms of known sequence-based information on the target itself and clearly homologous sequences. This was done using InterProScan[3]. The collected data was cross-referenced and

checked with the QuickGO[4] browser for the lowest compatible node in the GO tree. On some occasions, the knowledge of a structural template (identified via fold recognition) was used to infer possible molecular functions. The cellular component was also guessed with the aid of signal peptide predictions.

The tertiary structure prediction was based on results generated by our FOX server (see abstract by S. Toppo et al.). In homology modeling cases, the top scoring template was selected. For more difficult fold recognition targets, the results of the CAFASP meta server were cross-referenced with the FOX results. In particular, we took advantage of the extensive sequence space and back-validation data collected by FOX. Since the server stores the results for a large number of PSI-BLAST[5] searches in the sequence space, it is possible to highlight cases were solutions had been overlooked. In the most difficult cases, i.e. when PSI-BLAST found no other sequences, the choice was primarily based on secondary structure compatibility. For some difficult targets two or three different templates were selected for alignment and model generation.

The final choice for target-to-template alignment was based on the construction of raw 3D models and their evaluation with the Victor/FRST scoring function (see poster abstract by S.C.E. Tosatto for details). For both the target and template sequence a PSI-BLAST search (4 rounds) was started on the NR database clustered at 90% sequence identity to generate sequence profiles. Secondary structure was predicted for both sequences using PSIPRED[6]. This data was input into the profile/profile alignment program developed for the Arby server[7]. Rather than choosing a single alignment, four parameters (sequence and secondary structure weight, gap open and gap extension) were systematically modified to generate a total of 625 alignments. These were used to construct raw 3D models evaluated with Victor/FRST. The highest scoring alignment was chosen as the final alignment, with manual inspection of insertions and deletions limited to shifting gaps into loop regions, where appropriate. This alignment was submitted to the Homer server[8]. Conserved residue coordinates are copied, while indels are modeled using the fast divide & conquer method[9] and sidechains placed using SCWRL[10].

In cases where the homology to a known structure was evident, a more complex modeling strategy was used after alignment selection in order to increase local conformational sampling. The MOE modeling suite (Chemical Computing Group Inc.) is used to generate an ensemble of models for the target structure. The software generates independent models using a Boltzmann weighted randomized modeling procedure combined with a database search of fragments in the PDB[11] that cover insertions and deletions. Sidechains are modelled from a high-resolution rotamer library. The procedure ensures the construction of numerous variants of the model, which are evaluated with a residue packing quality function. The best models are selected for further automated refinement. The final model is chosen after visual inspection and evaluated with the Victor/FRST function from an ensemble of two to five locally minimized intermediate models.

In addition, another model (number 5) was submitted as the top scoring prediction for the Victor/FRST function for model quality estimation. This program had been entered for the MQAP (Model Quality Assessment Program) category in the CAFASP-4 experiment. Evaluation in this category yielded energy values for all models submitted to CAFASP by automated servers. It is therefore possible to evaluate how well this function performed in CASP-6 in assessing the quality of submitted server models.


1. Marchler-Bauer,A. et al. (2003) CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res.* **31**, 383-387.
2. Lexa,M., Valle,G. (2003) PRIMEX: rapid identification of oligonucleotide matches in whole genomes. *Bioinformatics*. **19**, 2486-2488.
3. Zdobnov,E.M., Apweiler,R. (2001) InterProScan--an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*. 17:847-848.
4. URL: *http://www.ebi.ac.uk/ego/*
5. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
6. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* ***292***, 195-202.
7. Von Ohsen,N., Sommer,I., Zimmer,R. and Lengauer,T. (2004). Arby: automatic protein structure prediction using profile-profile alignment and confidence measures. *Bioinformatics*. 20(14):2228-35.
8. URL: *http://protein.cribi.unipd.it/Homer/*
9. Tosatto,S.C.E., Bindewald,E., Hesser,J., Manner,R. (2002) A divide and conquer approach to fast loop modeling. *Protein Eng.* **15**, 279-286.
10. Canutescu,A.A., Shelenkov,A.A., & Dunbrack,R.L. (2003) A graph-theory algorithm for rapid protein side-chain prediction. *Protein Science*, **12,** 2001-2014.
11. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N., & Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res*, **28,** 235-242.

**CaspIta-FOX** (serv) - 315 models for 63 3D targets

## FOX (FOld eXtractor): a protein fold recognition method using iterative PSI-BLAST searches and structural alignments

P. Fontana[1], S.C.E. Tosatto[2], R. Velasco[1] G. Valle[2] and S. Toppo[3]
[1] *Istituto Agrario di San Michele all'Adige*
[2] *Dip. di Biologia & CRIBI Biotech Centre, Universita' di Padova*
[3] *Dip. di Chimica Biologica, Universita' di Padova*
stefano.toppo@unipd.it

We present a fold recognition method based on the combination of detailed sequence searches and structural information. Presently the protocol implements two different approaches to assign the most likely fold to the target protein sequence: the first is based on database secondary structure search and the second is based on iterative database sequence search.

In the first phase a secondary structure prediction of the target is performed based on the ConSSPred[1] protocol. This prediction is used to search for hits against a database of known secondary structures extracted from PDB (using DSSP). The search is based on a two-step strategy: the first step is based on a Smith-Waterman local secondary structure similarity search with a specific substitution matrix optimized for secondary structure alignment[2]. The second is based on a global alignment based on SSEA[3] (Secondary Structure Element Alignment), as implemented in our program MANIFOLD[4], to refine the score and the alignment itself in the region extracted from the first step. At the end of the first phase a list of hits that share a similar secondary structure topology with the target sequence is extracted.

The second phase is based on a modified protocol for scanning the sequence database called SENSER[5]. In the beginning of the second phase, BLASTP[6] is used to scan the target sequence against the NR database. These initial hits are clustered to reduce sequence bias and a seed alignment with 20 or fewer sequences generated. This step ensures that PSI-BLAST[7] can be jump-started with a more sensitive initial profile, increasing its sequence diversity. PSI-BLAST is run for four iterations (e-value inclusion threshold 10e-3) on the NR60 database of known sequences. NR60 is produced by applying the CD-HIT[8] algorithm to cluster the NR database at 60% sequence identity. Sequences producing NR60 hits with the query are assigned either to the significant sequence space (e-value <= 10e-3) or the trailing end (e-value <= 10) for further use. The profile is used to search the PDBAA database of sequences with known structure. If a significant PDBAA hit (e-value <= 10) is found, the protocol proceeds to the back-validation step (see below). If no significant hit is found, or the hit does not back-validate, a new PSI-BLAST search, using the above "4+1" protocol on NR and PDBAA, is started for the highest ranking sequence (i.e. lowest e-value) in the significant sequence space. Sequences from NR60 matching the query are also assigned to either the significant sequence space or the trailing end. Significant PDBAA hits are again submitted to back-validation. If no significant PDBAA hit is recorded and the significant sequence space has been exhausted, then the protocol uses the trailing end sequences as additional starting points for PSI-BLAST searches. In contrast to previous sequences, which were assumed to be similar enough to the target to imply homology, these sequences are submitted to back-validation before proceeding to the "4+1" PSI-BLAST protocol. The back-validation step consists in using PSI-BLAST to find the target starting from a different query sequence, found as described above. I.e. due to the asymmetric nature of PSI-BLAST, if sequence A finds sequence B it is not always the case that B also finds A. Sequences that back-validate are more likely to be correct hits. Once a sequence from PDBAA back-validates and its secondary structure is compatible with the one of the target sequence as found in the first phase, the protocol builds a target to template alignment and stops.

The procedure described so far serves to identify a template structure for the target sequence. In order to produce an accurate alignment, a profile-profile alignment approach has been used. The method is based on a program developed for the Arby server[9] which uses information from secondary structure predictions and sequence profiles. Alignments are automatically generated by systematically testing 625 different parameter combinations involving the weigths given to sequence profile and secondary structure of both target and template. Five values of each parameter are tested and chosen from a reasonable range. Each target-template alignment is used to build a raw model whose quality is evaluated on the basis of its estimated quality (see abstract of S.C.E. Tosatto). The best scoring target-template alignment is chosen to build and refine the final model.

The final model is generated using the package HOMER (http://protein.cribi.unipd.it/Homer). This involves the following steps. First a raw model of the conserved parts is constructed from the template. The conserved backbone 3D coordinates are copied and missing side chains placed with SCWRL[10]. Insertions and deletions are reconstructed using an enhanced version of the fast divide & conquer loop modeling method[11]. An experimental version of the FOX server is available at the following website address http://protein.cribi.unipd.it/fox.

1. Albrecht,M., Tosatto,S.C.E., Lengauer,T. and Valle,G. (2003). Simple consensus procedures are effective and sufficient in secondary structure prediction. *Protein Engineering* **16**, 459-462.
2. Wallqvist,A., Fukunishi,Y., Murphy,L.R., Fadel,A. and Levy,R.M. (2000). Iterative sequence/secondary structure search for protein homologs: comparison with amino acid sequence alignments and application to fold recognition in genome databases. *Bioinformatics* **16**(11), 988-1002.
3. Fontana,P., Bindewald,E., Toppo,S., Velasco,R., Valle,G. and Tosatto,S.C. (2004). The SSEA server for protein secondary structure alignment. *Bioinformatics*. Sep 3 [Epub ahead of print]
4. Bindewald,E., Cestaro,A., Hesser,J., Heiler,M. and Tosatto,S.C.E. (2003). MANIFOLD: Protein fold recognition based on secondary structure, sequence similarity and enzyme classification *Protein Engineering* **16**(11), 785-789.
5. Koretke,K.K., Russell,R.B. and Lupas,A.N. (2002). Fold recognition without folds. *Protein Science* **11**, 1575-1579.
6. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990). Basic Local Alignment Search Tool. *J. Mol. Biol.* **215**, 403-410.
7. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acid. Res.* **25**, 3389-3402.
8. Li,W., Jaroszewski,L. and Godzik,A. (2002). Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics* **18**, 77-82.
9. Von Ohsen,N., Sommer,I., Zimmer,R. and Lengauer,T. (2004). Arby: automatic protein structure prediction using profile-profile alignment and confidence measures. *Bioinformatics*. **20**(14), 2228-35
10. Canutescu,A.A., Shelenkov,A.A. and Dunbrack,R.L.Jr. (2003). A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci.* **12**(9), 2001-14.
11. Tosatto,S.C.E., Bindewald,E., Hesser,J., Manner,R. (2002) A divide and conquer approach to fast loop modeling. *Protein Eng.* **15**(4), 279-86.

**CBRC-3D** - 319 models for 64 3D / 22 FN targets

## Comparative modeling and fold recognition using FORTE series

K. Tomii, T. Hirokawa, and C. Motono
*Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, 2-43 Aomi, Koto-ku, Tokyo, Japan*
k-tomii@aist.go.jp

For fold recognition, profile-profile comparison is a powerful method to identify the structural similarity of two proteins that are compared. The method is also highly effective in improving sequence-structure alignments, even in comparative modeling. Recently, Wang and Dunbrack Jr.[1] performed a large series of benchmark tests using several profile-profile comparison methods. They suggested the effectiveness of deriving sequence-structure alignments from different protocols. We developed several descendants of our profile-profile comparison method[2] to make use of known structural information for protein structure prediction. Our prediction strategy in CASP6 is simple. For every target of CASP6, we have derived target-template alignments from several different protocols of profile-profile comparisons. We then constructed and exhaustively evaluated 3D models based on those alignments. Then we selected proper model(s) among them. We have specifically addressed the validation of our simple approach for protein structure prediction through CASP6. Our team was able to improve the selection of good models according to the fold recognition result in CASP5[3]. Consequently, we applied a more stringent method for 3D-model evaluation this time.

We devised three automated servers for fold recognition to investigate the possibilities of different profile-profile comparison protocols: FORTE1[2], FORTE2, and FORTE1T. The first, FORTE1, is the simple profile-profile comparison technique that is also used in CASP5. FORTE2 performs the same protocol as FORTE1 for profile-profile comparison using enhanced profiles. FORTE1T is a somewhat novel procedure of profile-profile comparison. All three of their servers were involved independently in CASP6. Aside from those three servers, we have developed and employed two systems that are also based on a profile-profile comparison method. One is a system, FORTE-H, that has hybrid profiles which contain sequence and secondary structure information. This system was inspired by a paper of Tang *et al.*,[4] but the generation and formulation of profiles differ slightly from their reported method. The other system, FORTE-SS, was developed for local profile-profile alignments. This

system was used mainly for exploring local structural templates when we failed to find global structural similarity of targets to known protein structures.

Modeling of a target protein based on the target-template alignments from four or five FORTE servers consists of two modeling process: (1) preliminary 3D model generation for master template selection, and (2) refinement of target-template alignments and reconstruction of accurate 3D models. We have controlled the variety or closeness of sequences that are included in profiles to be compared. Thereby, we have refined the alignments. Both processes, mainly used a molecular modeling program, MODELLER[5], for generating the 10 full-atom models for a target-template alignment. All of the generated models were evaluated based on a structural quality score (q-score) calculated using Verify3D[6] and Prosa2003[7] programs. This combination scheme for structural evaluation is more stringent than using a single evaluation method in our model selection.

Exhaustive modeling was performed in the preliminary 3D-model generation process. It used available templates (maximum 100 templates each for FORTE servers). The number of applied templates was reduced only for CM and easy FR targets that had promising templates with extremely high FORTE Z-scores. Acceptable 3D models from all candidates for the master templates in the next stage were estimated using their q-scores. The refinement process reconstructed 3D models of targets using a multi-template modeling approach with a master template and its structural neighbors, which were collected from a VAST[8] server after refinement of the target-template alignment if the neighbors were available. Secondary structure prediction and expected residue-residue contact information was included in the MODELLER restraint parameters to refine the local structures. We selected final models by their q-scores and human intervention when related knowledge was available from literature or other bioinformatics analysis results.

In addition to the procedures stated above, the use of information of S-S bonds provided additional evidence to choose templates for some cases. For possible FR/NF targets, some local structures or segments were constructed and validated using a library[9] of sequence-structure relationships derived from a known structure database. Some CM models were refined using their MD simulations. Functional predictions of targets were produced by observing motif conservation, employing knowledge from literature, using evolutionary trace method, and using human intervention of sequence conservation.


1. Wang,G. & Dunbrack,R.L. Jr. (2004) Scoring profile-to-profile sequence alignments. *Protein Sci.* ***13***, 1612-1626.
2. Tomii,K. & Akiyama,Y. (2004). FORTE: a profile-profile comparison tool for protein fold recognition. *Bioinformatics* ***20***, 594-595.
3. Kinch,L.N., Wrabl,J.O., Krishna,S.S., Majumdar,I., Sadreyev,R.I., Qi,Y., Pei,J., Cheng,H. & Grishin,N.V. (2003). CASP5 assessment of fold recognition target predictions. *Proteins*. ***53,*** 395-409.
4. Tang,C.L., Xie,L., Koh,I.Y., Posy,S., Alexov,E. & Honig,B. (2003). On the role of structural information in remote homology detection and sequence alignment: new methods using hybrid sequence profiles. *J. Mol. Biol.* ***334***, 1043-1062.
5. Sali,A. & Blundell,T.L. (1993). Comparative protein modeling by satisfaction of spatial restraints. *J. Mol. Biol.* ***234***, 779-815.
6. LÜthy,R., Bowie,J.U. & Eisenberg,D. (1992). Assessment of protein models with three-dimensional profiles. *Nature* ***356***, 83-85.
7. Sippl,M.J. (1993). Recognition of errors in three-dimensional structures of proteins. *Proteins* ***17***, 355-362.
8. Gibrat,J.F. Madej,T. & Bryant,S.H. (1996). Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.* ***6***, 377-385.
9. Tomii,K. (unpublished)


## **CBRC-DR** - 184 models for / 64 DP / 64 DR targets

### Prediction of disordered coil regions in proteins by threading and secondary structure prediction


T. Noguchi[1], S. Hirose[2], K. Shimizu [1,3] and K. Tomii[1]

*[1] – Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, Japan [2] –Pharma Design, Inc., Japan*

*[3]- Graduate School of Science & Engineering, Waseda University, Japan*

noguchi-tamotsu@aist.go.jp


We predicted structurally disordered coils in protein sequences using a protocol based on the following three steps: 1) We identified putative coil regions using threading methods (FORTE1[1], Superfamily[2] and SAM-T99[3]) combined and/or complemented with secondary structure predictions (PSIPRED[4], PHD[5], Jpred[6], Sspro[7], Prof[8] and SAM-T99); 2) We calculated the disorder propensity of the putative loop regions identified above. 3) Finally, we checked that the above predicted disordered regions were not inter-domain regions using domain linker prediction programs (DLP[9] and DomCut[10]). The predictions were performed at the META-PP meta-server[11], except for FORTE1, DLP and DomCut.

In step 1, loop regions were determined using homology modeling with FORTE1 only when the target's scores were larger than 10. In this case, homology modeling was reliable and coil regions of a query sequence were assigned by aligning the target protein to the template protein sequence of FORTE1. For targets whose fold identified with scores between 5 and 10 (5<=FORTE1<10), the secondary structures were not reliably determined by single threading method (FORTE1). Thus, we identified the coil regions of a target sequence by a consensus alignment on the template structure by three threading methods. Furthermore, when the template structures differed among the 3 threading methods, the alignment on the template with the highest FORTE1 score was used. Consensus secondary structure predictions were used to identify coils in regions, which were not assigned by threading. We prioritized predictions of PSIPRED, when no consensus secondary structure prediction was obtained. For sequences defined as new fold with FORTE1 (score<5.0), coil regions of the target sequence were assigned by the consensus secondary structure predictions.

In step 2, the disorder propensity for amino acid type was calculated using a non-redundant (sequence identity less than 30% and sequence length of a disorder region more than 5) PDB chain set compiled by PDB-REPRDB[12]. For the 700 representative chains the disordered regions were identified in the same manner as for DISOPRED[13], namely by comparing the SEQRES and the ATOM records in the PDB file and identifying the residues for which alpha carbon atoms coordinates are missing. Three sets of propensity scores were calculated for each sequence by dividing the sequence into an N-terminal, C-terminal and a central region.

We predicted disordered loop regions in proteins using the propensity and the loop regions as defined above, and according to the following criteria. All coil regions with three or more consecutive amino acids with high propensity and with an average propensity greater than 1.2 were predicted to be structurally disordered.

In the last step, we used two domain linker prediction methods to verify that the predicted disordered regions do not belong to inter-domain regions. We prioritized predictions of DLP, when no consensus domain linker prediction was obtained.


1. Tomii,K. & Akiyama,Y. (2004). FORTE: a profile-profile comparison tool for protein fold recognition. *Bioinformatics* **20**, 594-595.
2. Madera,M., Vogel,C., Kummerfeld,S.K., Chothia,C. & Gough,J. (2004) The SUPERFAMILY database in 2004: additions and improvements. *Nucl. Acids Res.* **32**, D235-D239.
3. Karplus,K. Barrett,C &Hughey,R. (1998). Hidden Markov Models for Detecting Remote Protein Homologies. *Bioinformatics* **14**, 846-856.
4. McGuffin,L.J., Bryson,K, Jones,D.T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics* **16**,404-5.
5. Rost,B. & C Sander,C. (1994). Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins* **19**, 55-77.
6. Cuff,J.A. & Barton,G.J. (1999). Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins* **34**, 508-519.
7. Pollastri,G., Przybylski,D., Rost,B. & Baldi,P. (2002) Improving the Prediction of Protein Secondary Structure in Three and Eight Classes Using Recurrent Neural Networks and Profiles. *Proteins* **47**, 228-235.
8. Ouali,M., & King,R.D. (2000) Cascaded multiple classifiers for secondary structure prediction. *Prot. Sci*. **9**, 1162-1176.
9. Miyazaki,S., Kuroda,Y. & Yokoyama,S. (2002). Characterization and prediction of linker sequences of multi-domain proteins by a neural network. *Journal of Structural and Functional Genomics* **2**, 37-51.
10. Suyama,M. & Ohara,O. (2003). DomCut: prediction of inter-domain linker regions in amino acid sequences. *Bioinformatics* **19**, 673-674.
11. Rost,B. & Liu,J. (2003). The Predict Protein Server. *Nucleic Acids Res*., **31**, 3300-3304.
12. Noguchi,T., Matsuda,H. & Akiyama,Y. (2003). PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB) in 2003. *Nucleic Acids Res*. **31**, 492-493.
13. Jones,D.T. & Ward,J.J. (2003). Prediction of disordered regions in protein from position specific matrices. *Proteins* **53**, 573-578.


## **CBRC-DR-SVM** - 108 models for 55 DR targets

### Predicting protein disordered regions using SVMs


K. Shimizu [1,3], S. Hirose[2] and T. Noguchi[1]

*[1] – Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, Japan ; [2] – Pharma Design, Inc., Japan; [3] – Graduate School of Science & Engineering, Waseda University, Japan*

kana@muraoka.info.waseda.ac.jp


We predicted protein disordered regions using a machine learning approach. This method has three steps. In the first step, protein-secondary-structures are predicted via PSI-PREPD[1]. In the next step, inputted sequences are divided into sliding-windows of size *m* (If the window is on terminal areas, *m* =7. If not,

*m*=15.). Finally, sequence features are extracted from the windows, and then each window is classified as an order or a disorder using Support Vector Machine (SVM)s, which is a powerful classification algorithm.

We prepared 834 attributes as input feature vectors for SVMs. These attributes included (A) 20 amino-acid compositions, (B) flexibility, (C) 10 compositions based on the physico-chemical characteristics of amino acids, (D) 400 compositions of adjacent amino acids, (E) 400 compositions of two amino acids with one residue between the two, and (F) results from protein-secondary-structure predictions which have possibility scores of 3 attributes, Helix, Beta-sheet, and Coil. Flexibility was calculated based on the score of normalized flexibility parameters[4], and all the physico-chemical characteristics of amino acids are listed in Table 1.

Since predictions depend on the positions of the windows, we learned and classified data separately according to the areas as follows. Basically, sequences are divided into an N-terminal area, a Central area and a C-terminal area. However, there is no concrete definition for "N-terminal area" or "C-terminal area", and windows on the boundaries include both central features and terminal features. Thus, we also learned and classified samples on the boundary areas, then integrated the results. We defined the N-terminal area as 0 to 10, the Central region as 10 to N-10, the C-terminal area as N-10 to N, and both boundary areas as 5 to 15 and N-15 to N-5. (The sequence length is N; where i means $i^{th}$ position from the N-terminal.)

Both the 693 non-redundant (sequence identity is less than 30% and the sequence length of disorder region is more than 5) PDB chains compiled by PDB-REPRDB[2] and the 135 chains that were listed in previous research[3], were used for learning process.

Table 1: Charactaristics of amino acids

| | |
|---|---|
| Hydrophilic | Aromatic |
| Hydrophobic | Aliphatic |
| Charged | Tiny |
| Plus | Small |
| Minus | Polar |

Basically, the method described above was used for all predictions, but different kernels for SVMs and different attributes were chosen depending on the situation. In model 1, the RBF kernel and 434 of the attributes (A+B+C+D+F) were used for all the predictions. In the N-terminal area of model 2, the polynomial kernel and 34 of the attributes (A+B+C+F) were used. In the Central area of model 2, the polynomial kernel and all of the 834 attributes were used. In the C-terminal area of model 2, the polynomial kernel and 434 of the attributes (A+B+C+D+F) were used.


1. McGuffin,L.J., Bryson,K, Jones,D.T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics* **16**,404-405.
2. Noguchi,T., Matsuda,H. & Akiyama,Y. (2003). PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB) in 2003. *Nucleic Acids Res*. **31**, 492-493.
3. Obradovic,Z., Peng,K., Vucetic,S., Radivojac,P., Brown,C.J., Dunker,A.K. (2003) Predicting intrinsic disorder from amino acid sequence. *Proteins:Struct. Funct. Genet.*, **53**, 566-572.
4. Vihinen,M., Torkkila,E., Riikonen,P. (1994) Accuracy of protein flexibility predictions. *Protein: Struct.Funct.Genet.*, **19**(2), 141-149.


## **CBSU** - 145 models for 64 3D targets

### Generation of protein structure models from fold-recognition and remote structural neighbors of templates


D. R. Ripoll and J. Pillardy

*Computational Biology Service Unit, Cornell Theory Center - Cornell University; Rhodes Hall Ithaca NY 14853-3801*

cbsu@tc.cornell.edu


We developed a protein structure prediction approach that was systematically applied to all the CASP6 targets. The principal source of structural information for each target was collected from the BIOINFO (3D-Jury)[1], LOOPP[2] and ROBETTA[3] servers. The templates used in the structure generation of our models were selected using the following conditions: (i) predictions from the three servers that consistently pointed to a structure (or domain of a structure) from PDB[4] (ii) if the servers provided predictions with low-level of confidence, only templates for which the secondary structure was highly consistent with that one predicted for the target sequence were further analyzed; (iii) whenever it was possible, structural alignments of the template structure with proteins sharing the fold but baring low sequence identity were constructed to identify the *essential* secondary structure elements and to determine the regions of high sequence variability. The Combinatorial Extension method[5] was used to obtain the corresponding structural neighbors having low sequence similarity (less than 30%) and relatively low (less than 5Å) $C^{\alpha}$ rms deviations with the template; (iv) for each template, attempts were made to improve the predicted sequence alignments provided by the servers by generating all-atoms 3D models where all essential elements associated with the template fold were present using the program MODELLER[6]. A set of rules were systematically

applied, e.g., (a) putative fragment deletion in the target sequence cannot eliminate a central strand of a β-sheet; (b) if an insertion falls inside an α-helical region, either the α-helical fragment is extended or the insertion is shifted toward the nearest loop region in the template fold. Otherwise, if the insertion falls in the middle of a β-strand, it is shifted toward the nearest loop region. In addition, a graphic program (DS-Modeling) was used to attempt to optimize further the alignment by using the hydrophilic/hydrophobic character of the residues.


1. Ginalski,K, Elofsson,A, Fischer,D, Rychlewski,L. (2003). 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics* **19**, 1015-1018; (http://bioinfo.pl/meta/)
2. Teodorescu,O, Galor,T, Pillardy,J, and Elber,R, (2004). Enriching the sequence substitution matrix by structural information. Prot., Struc, Funct. Bioinform. **54,** 41-48
3. Simons,K.T, Kooperberg,C., Huang,E., Baker,D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol*. **268**, 209-225.
4. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H, Shindyalov,I.N., Bourne,P.E. (2000). The Protein Data Bank. *Nucleic Acids Research* **28**, 235-242.
5. Shindyalov,I.N., Bourne,P.E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering* **11**, 739-747.
6. Šali,A., Blundell,T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779-815.


## **CHEN-WENDY** - 51 models for 23 3D targets

### Methods and algorithms in comparative protein modeling


J.L. Pellequer[1], G. Imbert[1], O.Pible[1], I. Vergely[1] and S-w. W. Chen[2]

*[1]-EA Valrhô – Centre de Marcoule – DSV/DIEP - Unit of post-genomic Biochemistry and Nuclear Toxicology. BP17171 – 30207 Bagnols sur Cèze – France; [2]-13 ave. de la Mayre – 30200 Bagnols sur Cèze – France*

cmft551@yahoo.com


Our comparative modeling approach is based on semi-automated prediction schemes with permanent user interventions. Putative template molecules were identified indulging in the CAFASP4 web server. We took advantage of the 3D Jury selection. Protein sequences of identified putative templates plus other homologous sequences were re-aligned using CLUSTALW/T-COFFEE[1]. The resulting multiple alignment was then manually refined with an in-house interactive tool to take into account the secondary structure of templates. Indel locations were refined by computational-graphics analysis of the three-dimensional structures of selected templates. In modelling CASP6 targets, a single template was used to build the target. We submitted multiple models corresponding to alternative template structures.

Side chain replacements and optimization were performed using our automatic program[2]. Replaced side chains were clustered and optimised in two steps: first, side-chain rotamers[3] were optimised at a cluster level, and second, the chi dihedral angles of each side chain were minimized at a residue level. Indels modeling were performed by optimizing backbone dihedral angles to close the loop gap. The coordinates of loop residues in deletions were taken from the template structure whereas residues in insertions were from our library. Side-chain conformations of loop residues were optimized according to the recipe of our side-chain positioning program. Additional minor refinements were performed by XPLOR energy-minimization in the CHARMM-22 force field[4]. To avoid over-minimization, the convergence criterion was set to between 1 and 4kcal/mol/Å while the Coulombic interaction was turned off for minimizing side-chain atoms. Each model was visually scrutinized to identify potential conflicts in side-chain conformations and to maximize side-chain-to-main-chain hydrogen bonds. Ranking models was performed by comparing the results of PROCHECK[5], PROSAII[6], VERIFY3D[7], ERRAT2.0[8], and hydrogen bonding (Pellequer & Chen, unpublished). When result is ambiguous, emphases were put on PROSAII.


1. Notredame,C., Higgins,D.G., & Heringa,J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**, 205-17.
2. Chen,S.-w.W. & Pellequer,J.L. (2004). Identification of functionally important residues in proteins using comparative models. *Curr Med Chem* **11**, 595-605.
3. Tufféry,P., Etchebest,C., Hazout,S. & Lavery,R. (1991). A new approach to the rapid determination of protein side chain conformations. *J Biomol Struct Dynam* **8**, 1267-1289.
4. MacKerell,A.D., Bashford,D., Bellott,M., Dunbrack,R.L.,Jr., Evanseck,J.D., Field,M.J. et al., (1998) All-atom empirical potential for

molecular modeling and dynamics studies of proteins. *J Phys Chem* **B102**, 3586-3616.
5. Laskowski,R.A., MacArthur,M.W., Moss,D.S., & Thornton,J.M. (1993). PROCHECK: A program to check the stereochemical quality of protein structures. *J Appl Cryst* **26**, 283-291.
6. Sippl, M.J. (1993). Recognition of errors in three-dimensional structures of proteins. *Proteins* **17**, 355-362.
7. Lüthy,R., Bowie,J.U., & Eisenberg,D. (1992). Assessment of protein models with three-dimensional profiles. *Nature* **356**, 83-85.
8. Colovos,C. & Yeates,T.O. (1993). Verification of protein structures: pattern of nonbonded atomic interactions. *Protein Sci* **2**, 1511-1519.

## **CHIMERA** - 65 models for 64 3D targets

### A versatile web user interface system for highly accurate protein structure prediction: SKE (Sophia-kai-Ergon) CHIMERA

M. Takeda-Shitaka, G. Terashi, D. Takaya, K. Kanou,
M. Iwadate and H. Umeyama
*Kitasato University*
shitakam@pharm.kitasato-u.ac.jp

Methods

Our laboratory registered group CHIMERA in CASP6 and groups FAMS and FAMD in CAFASP4. CHIMERA is a partially automatic modeling system that enables human intervention at necessary stages. Procedures of groups FAMS and FAMD, fully automated modeling servers, are very important and essential for large-scale genome modeling. In many cases, however, the procedures using human intervention are more accurate than fully automated modeling procedures. In previous CASP5, the results demonstrated that group CHIMERA constructed more accurate models than FAMS and FAMD did. In CASP6, we developed SKE CHIMERA, a web user interface system for highly accurate protein structure prediction based on the CHIMERA modeling system. This system enables human intervention at necessary stages easily.

The modeling procedure is 1) selection of reference proteins, 2) alignments, and 3) construction of model structures. Accuracy of the models depends on selection of reference proteins and on generating alignments. If reference proteins and alignments are wrong, model structures become wrong even though the modeling software is reliable. Therefore, we laid emphasis on these steps. These steps are based on the results of eight kinds of methods, BLAST, PSI-BLAST, PSF-BLAST, RPS-BLAST, IMPALA, FASTA, Pfam and PRED-FASTA (see abstracts of groups FAMS and FAMD). In high homology cases, we selected reference proteins from these results according to the secondary structure predictions. In low homology cases, we considered the reference proteins shown by automatic fold recognition servers in addition to the reference proteins shown by eight programs.

We generate alignments taking biologically important region, secondary structure predictions, homology, hydrophobic core etc. into consideration. Multiple templates are used when possible.

Based on the alignments, we constructed model structures using CHIMERA modeling system or FAMS modeling system. This step was automatic in most targets.

Results and Discussion

At present, the X-ray structures of 23 target domains have been released. Then we compared our models of groups CHIMERA, FAMS and FAMD with the corresponding X-ray structures, and calculated GST_TSs. As a result, 18 models of group CHIMERA were more accurate than those of FAMS and FAMD. 5 models are almost equal. These results demonstrated that our procedure that enables human intervention at necessary stages improves the model quality. In the post-genomic era, our highly accurate protein structure prediction system is essential for investigation of protein function, structure based drug design etc.

1. Takeda-Shitaka,M., Takaya,D., Chiba,C., Tanaka,H. and Umeyama,H. (2004) Protein structure prediction in structure based drug design. *Curr. Med. Chem.* **11**, 551-558.
2. Yoneda,T., Komooka,H. and Umeyama,H. (1997) A computer modeling study of the interaction between tissue factor pathway inhibitor and blood coagulation factor Xa. *J. Protein Chem.* **16**, 597-605.
3. Ogata,K. and Umeyama,H. (2000) An automatic homology modeling method consisting of database searches and simulated annealing. *J. Mol. Graphics Mod*. **18**, 258-272.

**Cinzia** - 48 models for 48 FN targets

## The melting pot of tools for function prediction

R. Calabrese[1], P. Fariselli[1], I. Rossi[1,2] and R. Casadio[1]
*[1] – Departement of Biology University of Bologna ,Via Irnerio 42 40126 Bologna (It), [2] – BioDec s.r.l , Via Fanin 48 40127 Bologna (It)*
casadio@alma.unibo.it

Our procedure for target function prediction is essentially based on the application of an ensemble of web tools, including ours, specifically suited for database mining, sequence alignment, sequence comparison, post translational modification, and protein structure prediction.

In our procedure, the first step consists in retrieving and collecting all information of interest for function prediction from a number of databases such as: Pfam[1], InterPro[2], Pir[3], Prosite[4] , SwissProt[5] , PDB[6] and others.

Then, when necessary, in order to confirm or add new features to the annotation of the target we proceeded as follows: first the target was aligned with PsiBlast[7] towards the june/04 nr release to retrieve a multiple sequence alignment; the formatted output was then routinely visualized with Jalview[8] to find well conserved region. Blocks of aligned sequences were selected for further refinement and by means of the PRATT[9] program a consensus pattern therefrom derived was used to scan the Prosite database for annotated sequences similar to our target.

Along with the above procedure we also used the PSIBlast derived profile (PSSM) to search the PDB for homologues of known structure and function

For annotating post translational modifications, and phosphorylation, we have merged outputs deriving from two servers available online ( http://www.cbs.dtu.dk/servicesProtFun;http://www.scansite.mit.edu/motifscan_seq.phtml). In this case only the prediction with the maximal score by both servers were retained.

For predicting the presence of disulfide bridges, we have used a neural network based tool developed by our research group, called CYSPRED that predicts the cysteine bonding state (http://www.biocomp.unibo.it)

1. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L.L., Studholme,D.J., Yates,C. & Eddy,S.R. (2004). The Pfam Protein Families Database. *Nucleic Acids Res.* **32**, 138-141.
2. Mulder,N.J., Apweiler,., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P., Bucher,P., Copley,R.R., Courcelle,E., Das,U., Durbin,R., Falquet,L., Fleischmann,W., Griffiths-Jones,S., Haft,D., Harte,N., Hulo,N., Kahn,D., Kanapin,A., Krestyaninova,M., Lopez,R., Letunic,I., Lonsdale,D., Silventoinen,V., Orchard,S.E., Pagni,M., Peyruc,D., Ponting,C.P., Selengut,J.D., Servant,F., Sigrist,C.J.A., Vaughan,R, Zdobnov,E.M. (2003). The InterPro Database, 2003 brings increased coverage and new features . *Nucleic Acids Res.* **31,** 315-318.
3. Wu,C.H., Huang,H., Nikolskaya,A., Hu,Z., Yeh,L.S., Barker,W.C. (2004). The iProClass Integrated database for protein functional analysis. *Computational Biology and Chemistry* **28,** 87-96.
4. Hulo,N., Sigrist,C.J.A., Le Saux,V., Langendijk-Genevaux,P.S., Bordoli,L., Gattiker,A., De Castro,E., Bucher,P., Bairoch,A. (2004). Recent improvements to the PROSITE database. *Nucl. Acids. Res.* **32**,134-137.
5. Bairoch,A., Boeckmann,B., Ferro,S., Gasteiger,E. (2004). Swiss-Prot: Juggling between evolution and stability. *Brief. Bioinform*. **5**,39-55
6. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig, H., Shindyalov,I.N., Bourne, P.E. (2000). The Protein Data Bank. *Nucleic Acids Research*. **28** , 235-242
7. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
8. Clamp,M., Cuff,J., Searle,S. & Barton, G. (1998). Jalview - Analysis and Manipulation of Multiple Sequence Alignments. (http://www.ebi.ac.uk/~michele/jalview/).
9. Jonassen,I. (1997). Efficient discovery of conserved patterns using a pattern graph. *Comput. Appl. Biosci.* **13(5)**, 509-522.
10. Martelli,P.L., Fariselli,P., Malaguti,L., Casadio,R. (2002). Prediction of the disulfide bonding state of cysteines in proteins with hidden neural networks. *Protein Eng.* **15**, 951-953.

**CLB3Group** - 268 models for 54 3D targets

# Predictor@home: a multiscale,distributed approach for protein structure prediction

C. An, M. Taufer and C.L. Brooks III
*The Scripps Research Institute,*
*10550 North Torrey Pines Road, La Jolla, CA 92037, USA*
brooks@scripps.edu

Motivation

In the previous CASP exercises we focused our efforts on addressing basic algorithmic and/or scientific questions related to the scoring of predicted protein structures and their refinement via all atom models. Retrospective analysis of our approaches and methods from these experiences suggested that when native-like protein conformations were sampled they could be identified with all atom physics-based force fields including implicit solvation[1]. During CASP6, we focused more directly on the question of conformational sampling, and whether, by augmentation of our earlier methods and algorithms by orders of magnitude more computational power, we could significantly improve our ability to predict protein structure. To achieve this objective we assembled a "structure prediction supercomputer" based on volunteered resources and a distributed computing platform using the world-wide-web in a project called Predictor@home.

Protocol for Protein Structure Prediction

Predictor@home approaches structure prediction through a multi-step pipeline that is similar to protocols that have led to successful prediction in the past[1]. In the first step of this pipeline, homology modeling and fold recognition templates are identified as significant hits from the BLAST and SAM-T02 servers. In addition, secondary structure is predicted by the PSIPRED server. The results from template recognition are used to generate restraints for aligned residues during lattice-based MFold simulations; untemplated regions are sampled by a Monte Carlo conformational search with the MONSSTER[2] force field using any available secondary structure information from PSIPRED. Secondary structure is the only information used to guide folding “new fold” prediction targets by MFold. In order to sample viable folded conformations, 5-10 thousand simulated annealing MFold tasks were distributed for each target, thereby increasing our sampling by 1-2.5 orders of magnitude over our past studies[1]. In the refinement step, each sampled structure is subjected to all-atom simulated annealing between 1000K and 300K using the molecular simulation package CHARMM and an intermediate accuracy all-atom force field. The lattice-based predictions provide inter-residue restraints implemented as NOE-like restraints based on side chain - side chain centers of mass contacts. Minimization is performed in the presence of the GBMV[3] solvent model to produce the final structure and energy value to be used in scoring. Scoring and ranking proceed via hierarchical clustering of the all-atom results based on the side chain contact-map.

The Architecture of Predictor@home

Predictor@home is built on top of the Berkeley Open Infrastructure for Network Computing (BOINC)[4]. BOINC is a well-known desktop grid framework that provides built-in support for distributed computing on heterogeneous PCs connected to Internet or Intranet networks. It currently supports a wide range of PC platforms (i.e., Linux, Windows, Mac, and Solaris). Protein structure prediction was achieved through two computationally intensive phases accomplished by two different codes:

1. MFold for protein structure assembly based on a low-resolution modeling method that uses a lattice representation;
2. CHARMM for protein refinement with an all-atom modeling method.

Predictor@home is a client-server based parallel computation paradigm. For each target, the server continuously generates MFold and CHARMM workunits (independent computations on a given target). The results from MFold are redirected by the server to CHARMM. Clients apply for computation and receive several workunits at a time. Client failures may occur and the returned results may be affected by hardware malfunctions or malicious attacks. Predictor@home addresses the integrity of the returned result using replicated computing and homogeneous redundancy (redundant instances of a computation are dispatched to numerically identical computers).

Over the course of the CASP6 season, we sampled over 430 thousand protein structures for 65 targets, each validated as the result of at least three replicas. In total nearly 7 thousand users registered for Predictor@home, with over 14 thousand machines.

1. Feig, M. & Brooks III, C. L. (2002). Evaluating CASP4 Predictions With Physical Energy Functions. *Proteins* **49**, 232-245.
2. Skolnick, J., Kolinski, A. & Ortiz, A. R. (1997). MONSSTER: A Method for Folding Globular Proteins with a Small Number of Distance Restraints. *J. Mol. Biol* **265**, 217-241.
3. Lee,M.S., Feig,M., Salsbury,F.R.,Jr. & Brooks,C.L.,III. (2003). New analytic approximation to the standard molecular volume definition and its application to generalized Born calculations. *J. Comput. Chem.* **24**, 1348-1356.

4. Anderson,D.P. (2003). BOINC: Berkeley Open Infrastructure for Network Computing. http:// boinc.berkeley.edu.

## **CMM/CIT/NIH** - 82 models for 42 3D targets

### Integrative refinement of homology models using colony energy approach with physical chemistry principles

Zhexin Xiang, Peter J. Steinbach
*Center for Molecular Modeling, Center for Information Technology, National Institutes of Health, Bethesda, Maryland 20892-5624*
xiangz@mail.nih.gov

Overview. Our participation in CASP6 tested the integration of knowledge-based and physics-based methods for protein structure refinement and the use of the so-called colony energy[1] to rank structures. We sought to automate the process of structure prediction, from template identification and alignment tuning to model refinement and verification, and subject the methodology to critical assessment. We did not try to identify the "best hit" if there was no clear agreement among servers. Instead, many template structures were considered for a given query sequence, the alignment for each template was refined using a genetic algorithm, and the submitted model was chosen based on its colony energy. We participated in the comparative modeling and fold recognition sections of the experiment, primarily using in-house software and methods, such as the JACKAL package (http://cmm.cit.nih.gov/~xiang/). JACKAL includes NEST (a new homology modeling program that is based on an artificial evolution method)[2], SCAP[3], and LOOPY[1] (a side-chain and loop prediction program), AUTOALIGN (a program to automatically tune a sequence alignment obtained from the CAFASP server), a physical-chemical based energy function to evaluate individual conformations, and the colony energy method to account for the clustering of conformations and energy space from fragment database and ab-initio sampling.

A three-step strategy was applied to fold recognition and to homology modeling: A) identify all possible templates, B) for each template, perform sequence-template alignment, C) for each model, refine any structurally variable (unconserved) regions and identify the best model.

(A) Template identification
CAFASP servers were used to identify as many prospective templates as possible. In the absence of a unanimous template identified by all servers, all possible hits were considered. For example, if multiple templates are identified but all servers point to the same structural family, all structures in the PDB from that family were used as possible templates. A family member unidentified by the servers was aligned with the query based on its structural alignment with its closest neighbor in the hit list. However, if more than 70% of servers agreed on one hit, and that hit also had the highest sequence similarity to the query among members of its structural family, then that particular hit was used as the sole template.

(B) Sequence Alignment
For each template identified in step (A), the alignments were ranked according to three factors: 1) sequence similarity to the query, 2) agreement with other servers, and 3) the physical-based energy of the model that corresponds to the alignment. We employed a "genetic-algorithm" approach to tune the alignment, and the best alignment was chosen from all the "offspring". Specifically, for each alignment obtained from a server, we produce large sets of candidate alignments by shuffling some of its alignment blocks with other alignments (either of the same template or of another template belonging to the structural family) reported by servers.

(C) Model building and Refinement
Because model building can be done rapidly using NEST, the ensemble of sequence alignments was readily converted to an ensemble of three-dimensional model structures. The models were clustered according to main-chain root mean-square deviation. The colony energy concept, developed initially for loop prediction, was then applied to energetically reward models that belong to large structural clusters in an attempt to approximately account for entropic effects. When used in loop modeling, the colony energy resulted in a smoothed energy surface [1].

All the alignments were converted to 3-dimentional models using the NEST program. This program builds and refines homology models using an artificial evolution method based on a single, composite or multiple templates. Given an alignment between a query sequence and a template, the alignment can be considered as a list of operations such as residue mutation, insertion or deletion. The algorithm always starts from the operation involving the smallest increase in an estimated physical-chemical energy. Each operation is followed be modest energy minimization to remove steric clashes. The final structure is then subjected to more thorough refinement. The structure-refinement module in NEST can refine the models in four levels: energy minimization of clashing atoms, refinement of insertion and deletion regions, refinement in all loop regions and refinement in all α/β regions. Refinement of helix or sheet regions

is done by a procedure similar to the loop refinement, but restraints are used to preserve main-chain hydrogen bonding.

For each model, unaligned regions corresponding to gaps in the sequence alignment were modeled using the LOOPY program. Two thousand initial conformations were randomly sampled and filtered against the consensus secondary-structure predictions from the CAFASP server. The 2000 conformations were then energy-minimized using our fast direct tweak method, and the 300 conformations of lowest energy were kept. An additional 300 were obtained from a fragment database using sequence similarity, secondary structure, and end-point geometry. The 600 conformations were subjected to additional energy minimization, and the conformation of lowest colony energy was selected. Structurally variable regions identified from multiple structure superimpositions were modeled similarly but the candidate conformations were restricted to those within 2 angstrom of the corresponding region in at least one of the PDB structures known for the structural family. Non-conserved side chains corresponding to mutations in the sequence alignment were modeled with the SCAP program. The conformation of a conserved side chains was unaltered unless its interactions with neighbors are strained (van der Waals energy >5 kcal/mol). A genetic algorithm was used to shuffle variable regions with other models; in variable (unconserved) regions, each refined model was married with other models to produce a large ensemble of offspring. The resultant candidates were then clustered and ranked using the colony energy. The model of lowest colony energy was inspected visually, and if satisfactory, submitted as our final prediction.


1. Xiang,Z., Soto,C. and Honig,B. (2002) Evaluating Conformational Free Energies: The Colony Energy and its Application to the Problem of Loop Prediction. *Proc. Natl. Acad. Sci. USA* **99**, 7432-7437.
2. Petrey,D., Xiang,X., Tang,C.L., Xie,L., Gimpelev,M., Mitors,T., Soto,C.S., Goldsmith-Fischman,S., Kernytsky,A., Schlessinger,A., Koh,I.Y.Y., Alexov,E. and Honig,B. Using Multiple Structure Alignments, Fast Model Building, and Energetic Analysis in Fold Recognition and Homology Modeling. *Proteins* **53,** 430-435.
3. Xiang,Z. and Honig,B. (2001) Extending the Accuracy Limits of Prediction for Side Chain Conformations. *J. Mol. Biol.* **311,** 421-430.


## **CoRind** - 18 models for 18 RR targets

### Quantitative measurement of covariation on an evolutionary tree with application to contact prediction


William J. Bruno and Aaron L. Halpern
*T-10, Los Alamos National Laboratory*
billb@lanl.gov


CoRind is a program that estimates the expected number of samples of the joint probability distribution for pairs of amino acids in two columns of a multiple alignment, given an evolutionary tree. It is entirely analogous to the Rind program which estimates independent samples for a single column[1]. Rind is unusual among evolutionary models in that it allows each column to have different amino acid frequencies. The implementation makes use of Felsenstein's[2] "pulley principle"(which is a form of dynamic programming) to propagate probabilities through the tree. A modified EM algorithm is used to converge to the maximum likelihood site-specific frequencies.

CoRind does not attempt to estimate frequencies for pairs of residues; rather it is designed to ask whether the evolution of a pair of sites can be adequately described by a model where the two sites evolve independently, with frequencies taken to be the products of the single site estimates. This is in contrast to the two-state likelihood ratio approach of Pollock et al.[3] (which offers a good review of other approaches to this problem).We note that although the CoRind model assumes the two sites evolve independently, the formula for the ancestral probabilities (used by the pulley principle) does not factor into separate expressions for the two sites. For example, if one sequence has the pair AA and another sequence has the pair CC, then the probability that their ancestor contributed the pair AC (viewed as a single letter in a 400 letter pair alphabet) to either sequence is zero (because neither has an AC) and this result cannot be obtained by treating both sites separately and combining the results. A previous prototype of Corind called Rind2 attempted to approximate the joint ancestral probabilities based the single site ancestral reconstructions, and the quality of the results seemed to vary greatly from one protein to the next.

Our "expected samples"are the same quantities that would be used to estimate the joint frequencies in the first iteration of the EM algorithm. We round off the expected number of samples to the nearest integer, and view the result as a contingency table to which Fisher's exact test[4] (FET) of independence may be applied. If all amino acids were to be used at both sites, the table would be 20x20, and the total number of counts would be at most the number of

sequences in the alignment (although usually substantially less due to evolutionary correlations). This is too large for exhaustive evaluation of FET, but it can be sampled numerically using a permutation test[5].

We apply a Bonferroni[6] correction for the number of pairs of sites that have enough variation (measured by calculating the autocovariation) to potentially covary at the required level of statistical significance. This correction is therefore expected to grow as the square of the sequence length. The number of samples required for one permutation test is proportional to this number, and since the number of tests to be done is also proportional to the square of sequence length, the total number of permutations needed scales as length to the fourth power. We improve on this in practice by automatically terminating permutation tests that we can be confident are not headed for a significant result.

The result of the permutation test followed by the Bonferroni multiple test correction is essentially an E-value for the observed covariation to have occurred by chance, assuming that the pair samples are independent (technically, that they obey the hypergeometric distribution, which fixes the marginal distributions for each site). For CASP a confidence value was computed by taking one minus the chance probability. A small pseudocount was applied to the raw permutation results so that a permutation test finding zero more strongly covarying (compared to the actual data) permutations out of one million yields a raw p-value of order $10^{-6}$ rather than zero.

Evolutionary trees were constructed by Weighbor[7] using the Rind[1] model. Weighbor was designed to be more robust than other fast tree reconstruction methods such as neighbor joining.

No where is any information used about scoring matrices or the evolutionary code. Each amino acid is treated as a unique letter in the alphabet. Gaps and X's are treated as unknown. The result is purely a test of covariation, without any prior assumptions about what the form of that covariation should be. This allows us to potentially find new patterns of protein evolution, but is bound to be a disadvantage from the point of view of pure 3D contact prediction.

Alignments were constructed using PSI-Blast[8] with default parameters and without any hand alignment editing. The only human intervention at this stage was determining how many iterations of PSI-Blast to allow. Alignments were not allowed to have more than 250 sequences, so PSI-Blast was stopped whenever this target was reached.

For 30 targets, no covariation was detected at the p<.5 level after Bonferroni correction, and no prediction was submitted. Proteins for which covariation was detected always had over 100 sequences in their multiple alignment; however, for three proteins with the largest allowed alignments of 250 sequences, no covariation was detected at this level of significance. Another 22 of the 76 targets expired before the software was implemented, or before it could complete its prediction (run time is usually only a few hours, but can be longer for long sequences with many pairs showing significant covariation). The remaining 24 targets were found to have significant covariation. For seven of these, the number of covarying pairs was suspiciously high, with some residues covarying with more than 10 other residues.

Our working hypothesis for the cause of such widespread covariation is that the functional pressure on the protein is different in different parts of the tree. This could be caused by an evolutionary change in function, such as a new substrate, or adaptation to some new condition, such as extreme temperature. In these cases human intervention was used to try to filter out sites that seem to correlate with any external evolutionary factor, as evidenced by strong covariation with an artificial (often binary) site representing which branch of the tree the sequence is from. For four of these cases, the result of the intervention was that after excluding sites that covary with the artificial site, and then excluding sites that covary with those, no covariation remained and no prediction was submitted. Homology models were consulted as a guide to this process. Specific information on how individual targets were handled was logged during the experiment at www.t10.lanl.gov/billb/corind.


1. Bruno,W.J. (1996) Modeling residue usage in aligned protein sequences via maximum likelihood. *Mol. Biol. Evol.* **13**, 1368-1374.
2. Felsenstein,J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**, 368-376.
3. Pollock,D.D., Taylor,W.R. and Goldman,N., (1999) Coevolving protein residues: maximum likelihood identification and relationship to structure. *J Mol. Biol.* **287**, 187-98.
4. Freeman,G.H., and Halton,J.H. (1951). Note on an exact treatment of contingency, goodness-of-fit, and other problems of significance. *Biometrika* **38**, 141-149.
5. Good.P. (2000) *Permutation tests: a practical guide to resampling methods for testing hypotheses*. Second edition, Springer Series in Statistics. Springer-Verlag, New York.
6. Bonferroni,C.E. Il calcolo delle assicurazioni su gruppi di teste. In *Studi in Onore del Professore Salvatore Ortu Carboni.* Rome: Italy, pp. 13-60, 1935.

7. Bruno,W.J., Socci,N.D., and Halpern,A.L. (2000) Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. *Mol. Biol. Evol.* **17**, 189-197.
8. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.



## CORNET: a server for the prediction of residue contacts in proteins

P. Fariselli[1], A. Valencia[2], and R. Casadio[1]
*1 - Iept. of Biology, University of Bologna via Irnerio 42 Bologna, Italy,*
*2 - Protein Design Group, CNB-CSIC Cantoblanco Madrid 28049, Spain*
piero@biocomp.unibo.it

We set up a web server (CORNET) for the prediction of contact maps of proteins using a neural network-based methods. Neural networks use an input encoding based on evolutionary information as derived by running PSI-BLAST[1] on a non redundant database of protein sequences.

CORNET uses the previously developed neural networks called NET[2]. NET has a single output neuron that codes for contact (output value close to 1) and non contact (output value close to 0). The hidden layer has 8 hidden neurons and the input consists of 1050 nodes representing the two possible pairing of two segments having a three-residue long window.

To be more detailed each residue pair in the protein sequence is coded as an input vector containing 210 elements (20x (20+1)/2), representing all the possible ordered pairs of residues (considering that each residue couple and its symmetric are coded in the same way). This is done in order to reduce the number of weight junctions. When single sequence is used, the input neuron coding for the ordered pair of amino acidic residues at positions i and j is set to 1, while the remaining 209 are set to 0. In order to take into account the sequence neighbours we use a 3-residue long input window, considering both parallel and anti-parallel pairing of the two segments centred at positions i and j, respectively. This leads to the coding of the couples formed by the residues in positions {i-1, j-1},{i, j},{i+1,j+1} (parallel pairing) and {i-1,j+1},{i,j},{i+1,j-1} (anti-parallel pairing) ending up with 5 possible combinations ({i-1,j-1}, {i,j},{i+1,j+1},{i-1,j+1},{i+1,j-1}) of the ordered couples. This is why this procedure requires 1050 (210x5) input neurons. Since we use multiple sequence information this binary input code is changed in a frequency-based one. This is done by considering the alignment from the corresponding PSI-BLAST[1] outputs and taking all the possible pairs generated by residues in positions i and j of the different aligned sequences. After normalization to the number of sequences, the frequencies of occurrence in the alignment of each couples is used in the corresponding position of the 210 element input vector

representing all the possible ordered pairs. By this, the 210 element vector may have more than one components activated.

To avoid contact overprediction, the predicted pairs are filtered taking into account the amount of contacts that each residue type can make (similarly to the procedure performed by Olmea and Valencia[3]). The filtering procedure is based on the occupancy data (or residue-coordination numbers) of each residue depending on its predicted secondary structure (we use a neural-network method whose overall 3-state accuracy reaches 0.74[4]). This value is derived from the set of protein structures of the data base and takes into account the secondary structure type. By this, the number of predicted contacts of a residue becomes a function of its structural environment. The occupancy can be therefore considered an estimate of the maximal number of contacts that each residue can make and is used to limit the number of contacts predicted for each residue.

On our dataset[2] we expect that NET accuracy (number of correct contact/ number of predicted contact) ranges from 0.23 to 0.07 depending on the protein lengths with an average of 0.16[2].


1. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
2. Fariselli,P., Olmea,O., Valencia,A., Casadio,R. (2001) Prediction of contact maps with neural networks and correlated mutations. *Prot. Engng.* ***14*** 835-843.
3. Olmea,O., Valencia,A. (1997). Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Fold Des.* **2**,S25-S32.
4. Jacoboni,I., Martelli,P.L., Fariselli,P., Compiani,M., Casadio,R. (2000). Predictions of protein segments with the same amino acid sequence and different secondary structure: a benchmark for predictive methods- *Proteins* **41**, 535-544.


## **Cracow.pl** - 140 models for 22 3D / 62 DR / 22 RR targets

## Early-stage folding in proteins – *in silico* model


M. Brylinski[1,2], L. Konieczny[3] and I. Roterman[2]

*[1] - Faculty of Chemistry, Jagiellonian University, Cracow, Poland,*
*[2] - Department of Bioinformatics and Telemedicine, Collegium Medicum, Jagiellonian University, Crakow, Poland; [3] - Institute of Biochemistry, Collegium Medicum – Jagiellonian University, Cracow, Poland*

brylinsk@chemia.uj.edu.pl, mbkoniec@cyf-kr.edu.pl, myroterm@cyf-kr.edu.pl


Verification of a model oriented on prediction of early-stage folding structures is the main goal of participation in CASP6. The model is based on:

1.The commonly accepted opinion that the early step of polypeptide chain folding is determined by the optimal conformation of the polypeptide backbone in the absence side chain-side chain interaction[1].

2.The ellipse-shaped conformational sub-space distinguishing on the Ramachandran map, optimal for the polypeptide chain backbone, linking all structurally important forms (helical and beta) [2,3].

3.The specific characteristics of the early-stage conformational sub-space expressing the balance between the amount of information stored in the amino acid sequence and the amount of information necessary to predict the conformation of early-stage folding[4].

4.The library we created expressing the relation between the sequence and early-stage folding conformation based on the known frequencies of $\varphi_e$, $\psi_e$ angles, which denote the $\varphi$, $\psi$ angles occurring in proteins after their transformation to the distinguished ellipse-shaped conformational sub-space.

The model oriented on early-stage folding prediction represents a universal approach which can be applied to any amino acid sequence independently on the length of the polypeptide chain. The model is believed to deliver the optimal starting structure for any procedures in an *ab initio* treatment.

A positive result is thus expected for secondary prediction, contact maps and assessing the degree of difficulty of structure prediction (although a high RMS-D value is possible). This is why the group is taking part in the following

categories: Order-Disorder Regions prediction and Residue-Residue separation distance prediction. Moreover, for targets less than 150 amino acid long, 3D atomic coordinates were predicted using procedure verified previously for BPTI[5], lysozyme[6], α chain of human hemoglobin[7] and ribonuclease[4].


1. Dobson,C.M. (2001) The structural basis of protein folding and its links with human disease. *Philos Trans R Soc Lond B Biol Sci*. **356**, 133-145.
2. Roterman,I. (1995) Modelling the optimal simulation path in the peptide chain folding--studies based on geometry of alanine heptapeptide. *J Theor Biol*. 177, 283-288.
3. Alonso,D.O. & Daggett,V. (1998) Molecular dynamics simulations of hydrophobic collapse of ubiquitin. *Protein Sci*. **7**, 860-874.
4. Jurkowski,W., Brylinski,M., Konieczny,L., Wisniowski,Z. & Roterman,I. (2004) Conformational subspace in simulation of early-stage protein folding. *Proteins* **55**, 115-127.
5. Brylinski,M., Jurkowski,W., Konieczny,L. & Roterman,I. (2004) Limited conformational space for early-stage protein folding simulation. *Bioinformatics* **20**, 199-205.
6. Jurkowski,W., Brylinski,M., Konieczny,L. & Roterman,I. (2004) Lysozyme folded in silico according to the limited conformational sub-space. *J Biomol Struct Dyn*. **22**, 149-158.
7. Brylinski,M., Jurkowski,W., Konieczny,L. & Roterman,I. (2004) Limitation of conformational space for proteins – early stage folding simulation of human α and β hemoglobin chain. *TASK Quarterly* **8**, 413-422.


## **cubic-chopper** (serv) - 40 models for 40 DP targets

### Automated domain boundary prediction using combination of sequence homology and neural network


J. Liu[1,2] and B. Rost[1,2]

*[1]– Dept. of Biochemistry & Molecular Biophysics, Columbia Univ.,*

*[2]– Center for Computational Biology and Bioinformatics, Columbia Univ.*

liu@cubic.bioc.columbia.edu


In the CASP 6 experiment, we tested an automated protein domain prediction server using combination of two previously published methods: CHOP[1;2], a method based on sequence similarity to known protein domains, and CHOPnet[3], a *de novo* domain prediction method using neural network.

The basic idea of CHOP was to identifies potential domain boundaries through hierarchical database searches beginning from very reliable experimental information (PDB[4]), proceeding to expert annotations of domain-like regions (Pfam-A[5]), and completing through cuts based on termini of known proteins (SWISS-PROT[6]). We have shown that CHOP can dissect over two thirds of all proteins from 62 proteomes, and the length distribution of fragments generated by CHOP resembles that of real protein domains.

CHOPnet was a method that predicts domain boundaries through a neural network using evolutionary information, predicted 1D structure (secondary structure, solvent accessibility), amino acid flexibility, and amino acid composition. The final predictions of domain boundaries resulted from post-processing the raw network output by removing noisy peaks. Cross-validation on proteins of known structure showed that CHOPnet correctly predicted the number of domains in 69% of all proteins. For 50% of the two-domain proteins, the centers of the predicted boundaries were within 20 residues of the true boundaries assigned from 3D structures.

Although CHOP can identify a considerable fraction of the structural domains reliably, it fails in the absence of sequence similarity to known protein domains. On the other hand, CHOPnet does not require the information from sequence homology, however, its accuracy dropped dramatically for proteins with more than two structural domains. In this context, we tested an automated method that combines the strength of CHOP and CHOPnet. The input sequence is first dissected with CHOP to identify domains that are similar to known structural domains from PDB or Pfam-A. Any remaining parts of the protein longer than 100aa are further processed by CHOPnet, in the hope that by pre-processing the sequence with CHOP, the number of unknown domain boundaries has been reduced, and CHOPnet can predict the remaining boundaries more accurately.

The method is available on-line at:
http://cubic.bioc.columbia.edu/services/CHOP/submit_casp.html


1. Liu,J. & Rost,B. (2004). CHOP: parsing proteins into structural domains. *Nucl. Acids Res.* **32**, W569-571.
2. Liu,J. & Rost,B. (2004). CHOP proteins into structural domain-like fragments. *Proteins* **55**, 678-688.
3. Liu,J. & Rost,B. (2004). Sequence-based prediction of protein domains. *Nucl. Acids Res.* **32**, 3522-3530.
4. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. & Bourne,P.E. (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235-242.

5. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. & Sonnhammer,E.L. (2002). The Pfam protein families database. *Nucl. Acids Res.* **30**, 276-280.
6. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I., Pilbout,S. & Schneider,M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucl. Acids Res.* **31**, 365-370.

## **DELCLAB** - 297 models for 62 3D targets

### Folding pattern recognition in proteins using spectral analysis methods

Carlos A. Del Carpio[1,2] & Jaime E. Barreda[2]
*New Industry Creation Hatchery Center, Tohoku University, Aoba-ku, Sendai 980-8579, Japan; [2] - Facultad de Farmacia y Bioquimica, Programa Profesional de Ingenieria Biotecnologica, Universidad Catolica de Santa Maria, Umacollo s/n, Arequipa, Peru*
cadcm@universia.edu.pe , jaimebarreda@universia.edu.pe

Divergence in sequence through evolution precludes sequence alignment based homology methodologies for protein folding prediction from detecting structural and folding similarities for distantly related protein. Homology coverage of actual data bases is also a factor playing a critical role in the performance of those methodologies, the factor being conspicuously apparent in what is called the twilight zone of sequence homology in which proteins of high degree of similarity in both biological function and structure are found but for which the amino acid sequence homology ranges from about 20% to less than 30%. In contrast to these methodologies a strategy is proposed here based on a different concept of sequence homology. This concept is derived from a periodicity analysis of the physicochemical properties of the residues constituting proteins primary structures. The analysis is performed using a front-end processing technique in automatic speech recognition by means of which the cepstrum (measure of the periodic wiggliness of a frequency response) is computed that leads to a spectral envelope that depicts the subtle periodicity in physicochemical characteristics of the sequence. Homology in sequences is then derived by alignment of spectral envelopes. Proteins sharing common folding patterns and biological function but low sequence homology can then be detected by the similarity in spectral dimension. The methodology applied to protein folding recognition underscores in many cases other methodologies in the twilight zone.

1. Del Carpio,C.A. Protein folding pattern recognition using signal processing theory. *Submitted for publication*.
2. Del Carpio,C.A. and Yoshimori,A. (2002) Fully automated protein tertiary structure prediction using Fourier transform spectral methods. Protein structure prediction: Bioinformatic approach, International University Line Publishers (IUL), 171-200.
3. Del Carpio C.A. and Carbajal J.C. (2002) Folding Pattern Recognition in Proteins Using Spectral Analysis Methods. *Genome Informatics* **13**, 163-172.

## **Distill** - 128 models for 64 3D / 64 RR targets

### Distill: fast, automated predictions of protein residue contacts and backbone coordinates by machine learning

G. Pollastri
*Computer Science Department, University College Dublin, Belfield, Dublin 4, Ireland*
gianluca.pollastri@ucd.ie

Distill is a fully automated system for ab initio prediction of draft protein structures. Distill has two main components: a set of predictors of protein features (secondary structure, relative solvent accessibility, residue contact maps, contact maps between secondary structure elements) based on machine learning techniques; an optimisation algorithm that searches the space of protein backbones under the guidance of a potential based on these protein features.

Protein secondary structure is predicted by Porter[1], an in-house system based on an ensemble of 45 bidirectional recurrent neural networks[2] with shortcut connections, accurate coding of input profiles obtained from multiple sequence alignments, second stage filtering and incorporation of long range information by a further layer of recurrent neural networks. Porter, tested by rigorous 5-fold cross-validation on a set of 2171 proteins, exceeds 79% correct classification on the “hard” CASP 3-class assignment, up to 81% on more lenient ones, making it one of the most accurate secondary structure predictors currently available.

Protein relative solvent accessibility, residue contact maps and maps of contacts between secondary structure elements are predicted by ensembles of recursive neural networks. These systems are recently trained, improved versions of the state-of-the-art ACCpro[3], CMAPpro[4] and CCMAPpro[5]. Residue contact maps submitted to CASP are further refined as follows: 10 backbones are reconstructed (see below) with a very short search (1,000 instead of 20,000 steps used in the full reconstruction); the contact maps of the 10 backbones obtained are averaged. This procedure is roughly as quick the initial contact map prediction (tens of second on a state-of-the art CPU) and cleans up the initial map of spurious, geometrically unrealisable contacts.

In the next stage, Distill reconstructs sets of backbone coordinates. The optimisation is carried out by minimising a simple potential function containing terms derived from the predicted features and terms representing geometrical constraints of the structure. Terms are present that penalise the violation of predicted residue contacts/non-contacts, predicted contacts/non-contacts between secondary structure elements, predicted strand locations, hard-core repulsion between amino acids, and virtual Cα-Cα bond lengths.
The actual minimisation is performed in 3 stages: 1) a set of initial structures is generated; 2) a search is performed from each initial guess, giving rise to a number of refined structures; 3) the final structures are ranked. In the initial guesses, helices and strands predicted by Porter are modelled, consecutive Cα atoms are set at a realistic distance (3.8±0.2Å), and virtual Cα angles are restricted to the 90º-180º interval. Each chain is grown from the N terminus to the C terminus by randomly selecting the next Cα with uniform distribution in the allowed space. A stochastic search from these initial guesses is performed by introducing perturbations in the structure similar to "crankshaft" moves[6], except that helices are treated as rigid "rods" and their core Cαs are never moved on their own. The search is carried out by simulated annealing with a linear schedule for the temperature. 20,000 moves of every non-helical Cα and helical termini are attempted for each search. 10 searches are run for each protein structure.

Finally, the 10 structures obtained are ranked. All mutual LCS at 1Å, 2Å, 4Å and 8Å are computed and each backbone is assigned a score equal to the sum of its LCS with the other backbones. The backbone with the highest score is selected and submitted to CASP. The rationale behind this ranking scheme is selecting the backbone containing most features common to most reconstructions.

Distill's modelling scheme is fast. On a cluster of 10 state-of-the-art PCs it can solve protein backbone coordinates on a genomic scale in the order of days.


1. Pollastri,G., McLysaght,A.. (2004) Porter: a new, accurate server for protein secondary structure prediction. *submitted*.
2. Baldi,P., Brunak, S., Frasconi, P., Soda.,G., and Pollastri, G. (1999) Exploiting the Past and the Future in Protein Secondary Structure Prediction. *Bioinformatics* **15**, 937-946.
3. Pollastri,G., Baldi,P., Fariselli,P., Casadio,R. (2002) Prediction of Coordination Number and Relative Solvent Accessibility in Proteins. *Proteins* **47**, 142-153.
4. Pollastri,G., Baldi,P. (2002) Prediction of Contact Maps by Recurrent Neural Network Architectures and Hidden Context Propagation from All Four Cardinal Corners. *Bioinformatics* **18**, S62-S70.
5. Baldi,P., Pollastri,G. (2003) The Principled Design of Large-Scale Recursive Neural Network Architectures -- DAG-RNNs and the Protein Structure Prediction Problem. *Journal of Machine Learning Research* **4**(Sep), 575-602.
6. Vendruscolo,M., Kussel,E., Domany, E. (1997) Recovery of protein structure from contact maps. *Folding and Design **2***, 295–306.


## **DomPRED** (serv) - 64 models for 64 DP targets
## **DomSSEA** (serv) - 64 models for 64 DP targets

### Protein domain prediction using the DomPred server


K. Bryson[1], L.J. McGuffin[1], R.L. Marsden[2] and D.T. Jones[1]
*[1] – Bioinformatics Unit, Department of Computer Science*
*[2] – Department of Biochemistry*
*University College London, Gower Street, London WC1E 6BT*
dtj@cs.ucl.ac.uk


The DomPred Server[1] contains our previously published method for domain prediction, DomSSEA[2], combined with a newly developed method called Domains Predicted from Sequence (DPS).

DomSSEA uses a fold recognition approach, based on aligning the PSIPRED[3] predicted secondary structure for the query sequence against the DSSP[4] assigned secondary structures of a fold library. It then transfers the SCOP[5] assigned domain structure from the best fold match to the query sequence.

DPS carries out a PSI-BLAST[6] search of the query sequence against a database consisting of NRDB90[7] augmented with sequences from Pfam-A[8]. Significant local alignment fragments are examined, and the total numbers of C- and N-terminals for the fragments are recorded for each residue position in the query sequence. These distributions are smoothed. They are then combined giving additional weight to positions which have high values for both the C- and N-terminals, since this provides more evidence for a domain boundary in which one conserved sequence region ends and another starts. The combined values are then turned into Z-scores by dividing throughout by the standard deviation over the entire query protein. A threshold is then applied to these z-score values in order to predict domain boundaries.

The DomSSEA method is most effective when the fold library contains a complete structural match to the query. Hence this approach bears some resemblance to remote homology detection or fold recognition. The DPS method makes no such use of complete structural matches, since the alignment would just have its N- and C-terminal positions lying close to the N- and C-terminals of the query sequence. Large scores close to the N- and C-terminal of the query sequence are simply excluded as end-effects when predicting domain boundaries, for obvious reasons. DPS relies more on the database containing sequences which have fragments that in combination reveal the domain structure of the protein, rather than a complete sequence with a structural match. Thus it can be seen that these two methods are based on largely orthogonal information, and hence they are particularly effective in combination.

Currently the results from the two methods are combined by the user. There are plans to form a consensus method, combining both of these approaches. Also we wish to have the server carry out an initial screening stage in which it detects obvious homologues to PDB structures and just reports back their domain structure. This will lead to a robust server which can deal with both easy and difficult cases.


1. http:/bioinf.cs.ucl.ac.uk/dompred/
2. Marsden,R.L., McGuffin,L.J. & Jones,D.T. (2002) Rapid protein domain assignment from amino acid sequence using predicted secondary structure. *Proteins Sci.* **11**, 2814-2824.
3. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* ***292***, 195-202.
4. Kabsch,W. & Sander,C. (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers* **22**, 2577-2637.
5. Andreeva,A., Howorth,D., Brenner,S.E., Hubbard,T.J.P., Chothia,C., & Murzin,A.G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic. Acids Res.* **32**, D226-D229.
6. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
7. Holm,L. & Sander,C. (1998) Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics* **14**, 423-429.
8. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S.,, Sonnhammer,E.L., Studholme,D.J., Yeats,C. & Eddy,S.R. (2004) The Pfam protein families database. *Nucleic Acids Res.* **32**, D138-D141.


## **Dopro** (serv) - 62 models for 62 DP targets

### Dopro: automatic protein domain structure prediction using a stochastic model for analyzing homology search results


N. von Öhsen[1], J. Apostolakis[2], R. Zimmer[2]

*[1] - Fraunhofer Institute for Scientific Computing and Algorithms (SCAI), Sankt Augustin [2]-Institute for Informatics, Ludwig-Maximilians-Universität München*

niklas.von-oehsen@scai.fraunhofer.de


The aim of the Dopro server is to combine the results from multiple runs of a homology search method with different subsequences of the target sequence into a model of the protein's domain structure. The prediction is calculated as the maximum likelihood estimate of the domain structure with respect to a probabilistic model that describes the probability of the set of homology search results given a specific domain structure. The method is restricted to predicting domains as intervals, i.e. no disrupted domains can be modeled.

The method starts by constructing a set of subsequences from the query sequence. A number of standard subsequences like first and second half and also the thirds are added to the set. Additionally, first guesses of correct domains are included. These are gathered by a PSI-Blast[1] search against the ProDom[2] database and by analysis of the predicted secondary structure: A secondary structure is predicted using PSIPRED[3] and segments of predicted loops are used as potential domain boundaries. Finally, the set of all subsequences is reduced to a reasonable size by removing subsequences that are highly similar or short.

For each subsequence, a multiple alignment is constructed by searching the NR database using PSI-Blast. A frequency profile is calculated from this multiple alignment using a slightly modified version of the Henikoff-Henikoff sequence-weighting algorithm[4]. This frequency profile is used for searching against a protein domain template database based on a 40% ASTRAL set containing frequency profiles for all domains[5]. The method used for the search is our profile-profile alignment method using the log average score[6-7]. The performance of the method has recently been independently assessed[8-9].

The top hits of the search are annotated with confidence measures developed for quantification of the reliability of the search results[10]. Also, the start and end points of the hits and the SCOP[11] classification of the template are recorded.

This data set of different subsequences with corresponding fold recognition results and start and end points of the target-template alignments is then further processed using the probabilistic model.

The probabilistic model describes how a protein domain structure (consisting of a list of intervals on the target sequence, each annotated with SCOP fold) produces the above determined set of fold recognition results. For such a domain structure model and for a subsequence of the target sequence, a set of probabilistic equations describes the probability of observing a top hit with a certain template fold, confidence value, and start/end point. Several steps are represented by these equations: First, one of the multiple domains is selected as source of the hit. A possible error in the fold recognition is also modeled depending on the above confidence measure. Finally, the distribution of start and end points is determined by allowing small variation around the intersection of query subsequence and selected domain.

In order to allow a rapid determination of the maximum likelihood estimate of the domain structure, the size of the model space is considerably reduced by allowing only the previously determined set of top hits as predicted domains. Thus, only all non-intersecting combinations of these have to be evaluated, for which an exhaustive search is feasible.

The infrastructure of the Dopro server is based on the same engine as the Arby protein structure prediction server[12], which is a Java based implementation of a data flow engine. The implementation features a fully parallel execution of the computationally involving tasks and was executed on a 12 CPU Sun system.

The Dopro server uses a combination of profile-profile homology detection methods and a stochastic description of their outcome to produce a prediction for the protein domain structure. Future developments might implement more sophisticated techniques like MCMC to extract the maximum likelihood estimate from an unrestricted search space.

*Acknowledgements*. The Dopro team wishes to thank E. Schrüfer, T. Mevissen and M. Hofmann from the Fraunhofer Institute SCAI and I. Sommer from the Max-Planck-Institute for Informatics for their support during the prediction season. This work was supported by the BMBF project "Development of Microbalance Array/Mass Spectrometry as a Tool for Functional Proteomics" (0312708).

1. Altschul,S.F. et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. ***25*** *(17)*, 3389-402.
2. Servant,F., Bru,C., Carre,S., Courcelle,E., Gouzy,J., Peyruc,D., Kahn,D. (2002) ProDom: Automated clustering of homologous domains. *Briefings in Bioinformatics* ***3***, 246-251.
3. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*. ***292*** *(2)*, 195-202.
4. Henikoff,S., Henikoff,J.G. (1994) Position-based sequence weights. *J Mol Biol*. ***243*** *(4)*, 574-8.
5. Chandonia,J.M. et al. (2002) ASTRAL compendium enhancements. *Nucleic Acids Res*. ***30*** *(1)*, 260-3.
6. von Öhsen,N. Zimmer,R. (2001) Improving profile-profile alignment via log average scoring. *Lecture Notes in Computer Science* ***2149***, 11-26.
7. von Öhsen,N., Sommer,I., Zimmer,R. (2003) Profile-Profile Alignment: A Powerful Tool For Protein Structure Prediction. in *Pac Symp Biocomput*.
8. Edgar,R., Sjölander,K. (2004) A comparison of scoring functions for protein sequence profile alignment. *Bioinformatics* ***20*** *(8)*, 1301-8.
9. Ohlson,T., Wallner,B., Elofsson,A., (2004) Profile-profile methods provide improved fold-recognition: a study of different profile-profile alignment methods. *Proteins* ***57*** *(1)*, 188-97.
10. Sommer,I. et al. (2002) Confidence measures for protein fold recognition. *Bioinformatics* ***18*** *(6)*, 802-12.
11. Lo Conte,L., Brenner,S., Hubbard,T.J., Chothia,C., Murzin,A.G. SCOP database in 2002: refinements accommodate structural genomics (2002) *Nucl. Acids. Res.* ***30***, 264-7.
12. von Öhsen,N., Sommer,I., Zimmer,R., Lengauer,T. (2004) Arby: automatic protein structure prediction using profile-profile alignment and confidence measures. *Bioinformatics* ***20*** *(14)*, 2228-35.

## **DRIP-PRED** (serv) - 64 models for 64 DR targets

### Order/disorder prediction with self organising maps

R.M. MacCallum
*Stockholm Bioinformatics Center, Stockholm University, Sweden.*
maccallr@sbc.su.se

We recently developed methods for sequence profile visualisation and contact map prediction[1] based on Kohonen's self organising map (SOM). The key issue in that work was the huge reduction in the dimensionality of sequence profile windows, which can typically contain over 300 values per sequence position, to a smaller more manageable 3D “colour space”. The resulting clusters/colours raised interesting questions about sequence-structure relationships in proteins,

particularly in strands and helices. It was also interesting to see how less frequent sequence patterns were treated by the SOM. Occasionally, unusual residue colouring would be seen in an unusual structure context, for example a buried helix, or certain loops. The SOMs for contact prediction were trained on sequences of known structure, but we were also compelled to train them on "all proteins" and see how local sequence window space looks when flattened into a SOM. One obvious question is: what it the overlap between "universal sequence window space" and "solved structure sequence window space"? One product of this research was the order/disorder predictor "DRIP-PRED".

Sequences from UniProt version 43 were made nonredundant using a crude single-pass Perl hashing approach; a sequence is discarded if it contains an exact 9 residue match with a previously encountered protein. The resulting set contains just 8003 proteins, but is sufficient for our purposes. Each sequence, of length *L* residues, is run through PSI-BLAST[3] using PSIPRED[2] version 2.3 scripts, in order to generate a ".mtx" text file containing the position specific scoring matrix of *L* columns by 21 rows. The rows correspond to the 20 amino acids and a mystery value, presumably related to indels. A total of *L* overlapping windows of width 15 are extracted from the matrix, using zeroes to pad at each end. In total, 3,084,456 windows (15 by 21 submatrices) are extracted from the 8003 proteins.

The 3,084,456 sequence profile windows are clustered/mapped into a SOM grid of 25 by 20 nodes. The training procedure is divided into 20 steps. For each step, 1/20 of the input data is sampled at random, and afterwards the training rate and neighbourhood radius are decreased linearly (starting at 0.05 and 12, respectively). After training, any (15 by 21) sequence profile window can be mapped to a discrete location on the SOM grid.

We next calculate the "hit frequencies" for each SOM node for different types of protein. This is simply the number of hits to each node divided by the total number of hits to the whole map. We have done these calculations for the representative UniProt sequences mentioned above and for a representative set of proteins of known structure (an ASTRAL 10% identity subset of SCOP 1.55). As can be seen at http://www.sbc.su.se/~maccallr/disorder/maps.html, there are regions of "UniProt space" which are essentially unpopulated by proteins of known structure. Sequence windows which map to these locations are not well represented in the PDB and therefore probably do not have an ordered 3D structure. This is the basis of the DRIP-PRED predictor. We quantify this by calculating the value log(UniProt/SCOP) for each position in the map (see web figure part C).

The target sequence is processed in the same way as the training data (see above) to produce a PSI-BLAST profile. Every window of 15 residues centred around residue *i* is mapped to a node on the UniProt SOM, and baseline disorder prediction score for this residue is taken from the corresponding position in the log matrix (web figure, part C). Note that at this point, the score is distributed around zero, with positive values indicating disorder. As in Jones and Ward[4], a confident secondary structure prediction (from PSIPRED[2]) suggests that there is some ordered structure, so we set the score to -0.5 when the numerical PSIPRED outputs (H for helix, E for strand, C for coil) for that residue satisfy the following:

$$H - (E+C) > 0.5 \quad \text{OR} \quad E - (H+C) > 0.5$$

The scores are then smoothed with four cycles of a ±1 residue window average, and then adjusted by +0.5, and finally capped to the range 0-1. Manual inspection of a few CASP5 targets suggested that a threshold of 0.5 was suitable to delineate between ordered and disordered (>0.5 is disordered). Note: *no other optimisation, training or evaluation was performed*. A web service is available at http://www.sbc.su.se/~maccallr/disorder/.


1. MacCallum,R.M. (2004) Striped sheets and protein contact prediction. *Bioinformatics*, **20** Suppl 1, I224-I231.
2. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195-202.
3. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
4. Jones,D.T. & Ward, J.J. (2003). Prediction of disordered regions in proteins from position specific score matrices. *Proteins*. **53** Suppl 6:573-8.


## **EBGM** - 12 models for 3 3D targets

### Sorting candidate models using a structural alphabet-amino acid sequence compatibility approach


P. Tufféry and A.C. Camproux

*Equipe de Bioinformatique Génomique et Moléculaire, INSERM E0346*
*Université Paris 7, case 7113, 2 place Jussieu, 75251 Paris cedex 05*
tuffery@ebgm.jussieu.fr


We have tested an approach that combines the encoding of protein structures into a sequence of letters of a Structural Alphabet (SA-sequence) and a

predictor of the SA-sequence from the Amino-Acid sequence (AA-sequence). We use a Hidden Markov Model derived structural alphabet of 27 states, as described in[1]. Such alphabet was shown to accurately describe the conformation of the proteins. This alphabet describes the conformation of overlapping fragments of 4-residue length and the way they can be interconnected.Using the Viterbi or the forward-backward algorithms, it is possible to identify the best sequence of letters that describe a protein structure in the terms of the structural alphabet.

From a collection of over 2675 non redundant proteins(less than 30% amino acid sequence identity), we have setup a method to predict the letters of the structural alphabet from its amino-acid sequence. It combines a Bayesian approach with the logic of the SA using the HMM procedure. Finally, we use such predictor to measure the compatibility between a given structure and its amino-acid sequence as follows:
(i) We encode the candidate models as SA-sequences as can be achieved at http://bioserv.rpbs.jussieu.fr
(ii) By constraining the prediction from the AA-sequence to fit the SA-sequence obtained from the 3D conformation of a model, it is possible to obtain an estimate of how the encoded structural model is compatible with the AA-sequence, in terms of likelihood. This likelihood of the AA sequence constrained by the SA-sequence can be understood as some kind of measure of the compatibility of the structure with its amino acid sequence.
(iii) Having computed the likelihoods associated with a series of candidate structural models, we select the models having the maximal likelihoods.

Here, we have applied such approach to targets labeled by the robetta server[2,3] as « cutpref », i.e. the lowest level of confidence of the robetta server, such as T0215, T0239, T0242, T0243.


1. Camproux,A.C., Gautier,R. & Tuffery,P. (2004) A Hidden Markov Model Derived Structural Alphabet for Proteins J. Mol. Biol. 339, 591-605.
2. Kim.,D.E., Chivian,D. & Baker,D. (2004) Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res. 32, Suppl.2, W526-531.
3. Chivian,D., Kim,D.E., Malmstrom,L., Bradley,P., Robertson,T., Murphy,P., Strauss,C.E.M., Bonneau,R., Rohl,C.A. & Baker, D. (2003) Automated prediction of CASP-5 structures using the Robetta server. *Proteins* **53**, Suppl 6, 524-533.


**Eidogen-SFST** (serv) - 64 models for 64 3D targets
**Eidogen-BNMX** (serv) - 64 models for 64 3D targets
**Eidogen-EXPM** (serv) - 64 models for 64 3D targets

## Automated structure modeling with Eidogen's suite of algorithms


A. Poleksic, J.F. Danzer and D.A. Debe
*Eidogen, Inc.*
aleksandar@eidogen.com


STRUCTFAST (Structure Realization Utilizing Cogent Tips From Aligned Structural Templates) is a novel profile-profile alignment algorithm uniquely capable of incorporating important information from a structural family directly into the dynamic programming process. Query sequence profiles are generated using a modified version of NCBI's PSI-BLAST algorithm[1]. A database of profiles for representatives from the PDB are generated in a similar manner, but are augmented with information from structure based alignments for the structural family. Each query sequence is then aligned and scored against the library of structural profiles. Statistical significance of alignment scores are assessed using a variant[2] of the island statistics method[3,4], so that the final E-value for every database hit accounts for the lengths and compositions of the sequences being compared.

The core alignment algorithm in all three of our automated servers is the same. SFST outputs the alignment with the overall best E-value. BNMX and EXPM go a step further to refine α-Carbon coordinates by using multiple PDB templates and the remaining backbone atoms are reconstructed from the α-Carbon coordinates[5]. The only difference between BNMX and EXPM is the choice of a few algorithm parameters, such as the score significance cutoffs and gap penalties.


1. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
2. Poleksic,A., Hambly,K., Danzer,J.F., Debe,D.A. Increased remote homology detection performance using a fast method for determining local alignment statistics, *unpublished*.
3. Olsen,R., Bundschuh,R. And Hwa,T. (1999) Rapid assessment of extremal statistics for gapped local alignment. In Lengauer,T., Schneider,R., Bork,P.,

Brutlag,D., Glasgow,J., Mewes,H.-W. and Zimmer,R. (eds), *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, AAAI Press, Menlo Park, CA, pp. 211-222.
4. Altschul,S.F., Bundschuh,R., Olsen,R., Hwa,T. (2001) The estimation of statistical parameters for local alignment score distributions. *Nucleic Acids Res.* **29**, 351-361.
5. Rey,A., Skolnick,J. (1992) Efficient Algorithm for the Reconstruction of a Protein backbone from the a-Carbon Coordinates. *J. Comput. Chem.* **13**, 443-456.

**EMBL_DisEMBL_coil** - 57 models for 57 DR targets
**EMBL_DisEMBL_hotloop** - 57 models for 57 DR targets
**EMBL_DisEMBL_rem465** - 57 models for 57 DR targets
**EMBL_GlobPlot** - 114 models for 57 DP / 57 DR targets

## Predictions of order/disorder in CASP6 using GlobPlot & DisEMBL

Rune Linding, Lars Juhl-Jensen, Toby Gibson, Robert B. Russell
*EMBL – Biocomputing, Heidelberg, Germany*
linding@embl.de

We applied two of our recently developed order/disorder predictors to CASP6 targets. For each target, we used default parameters for both GlobPlot (http://globplot.embl.de) & DisEMBL (http://dis.embl.de), augmented by some visual inspection based on observations about the target. GlobPlot predicts the tendency of segments within a protein sequence for order/globularity and disorder. It uses a simple set of parameters based on the tendency of amino acids to lie in structured or unstructured regions in known structures or proteins where tendency to order/disorder is known experimentally. DisEMBL addresses more explicitly the problem of identifying regions in a protein sequence likely to be disordred or absent in known structures. It uses a neural network trained on various measures of order/disorder extracted from known three-dimensional structures. Both GlobPlot and DisEMBL were developed for prediction of structural context of linear motifs as catalogued by ELM (http://elm.eu.org).

1. Linding R, Russell RB, Neduva V, Gibson TJ. (2003) GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res.* **31**(13), 3701-3708.
2. Linding,R., Jensen,L.J., Diella,F., Bork,P., Gibson,T.J., Russell,R.B. (2003) Protein disorder prediction: implications for structural proteomics. *Structure (Camb).* **11**(11), 1453-1459.

**FAMD** (serv) - 320 models for 64 3D targets

## Full automatic homology-modeling servers including wisdom and practice: SKE(Sophia Kai Ergon) FAMD

K. Kanou[1], M. Iwadate[1], G. Terashi[1], D. Takaya[1],
M. Takeda-Shitaka[1] and H. Umeyama[1]
[1] - *Department of Biomolecular Desig*
*School of Pharmaceutical Sciences, Kitasato University*
kanouk@pharm.kitasato-u.ac.jp

Selecting alignment

The aliment selection for constructing highly accurate protein models using homology modeling was described. 7 kinds of methods, BLAST[1], PSI-BLAST, PSF-BLAST, RPS-BLAST, IMPALA, FASTA and Pfam were executed for amino acid sequences of query proteins.

PSF-BLAST is PSI-BLAST whose profile sequence group of PSSM construction process is revised, and the selection criterion is E-value<=0.001 from template PDB sequence on PSI-BLAST search.

For selecting the best in 7 kinds of alignment methods, the score-function that was constructed by model length, homology% and degree of secondary structure agreement between PSI-PRED and STRIDE was defined:

$$score = f(k_i, Hom, Len, SS)$$

$Len$ is residue length of model protein. $Hom$ indicate homology % value, the ratio between the number of match residues and $Len$. $SS$ is so called Q3 value, degree of secondary structure agreement between PSI-PRED and STRIDE. $k_i$ are coefficients. The subscript number "i" indicate kind of alignment method, 0 is PSI-BLAST, 1 is BLAST, 2 is RPS-BLAST, 3 is Family-BLAST, 4 is IMPALA, 5 is FASTA, 7 is Pfam.

Selecting fragment

The fragment selection process of FAMS[2] was modified to select fragment of same protein family. Therefore selection criteria were RMSD of fitting, and degree of SCOP[3] ID agreement between template PDB and fragment.

Energy Minimization and Molecular Dynamics

After homology modeling, both of Energy Minimization and Molecular Dynamics are applied.

Results and Discussion

In 29 available CASP6 target structures, models were evaluated with GDT_TS. Number of targets that maximum GDT_TS with 7 kinds of methods were more than 30 was 19. The 19 targets approximately correspond to "PDB-Blast hits" targets in CAFASP website, and 10 targets approximately correspond to no "PDB-Blast hits" targets. Therefore, 10 targets were high-difficulty targets in 29 targets.

In 16 of 19 targets, the alignments that GTD_TS are more than 87% of maximum GDT_TS were selected with this score-function. High GDT_TS detection capability with a simple score-function was indicated.

The server was same to FAMS server that PRED-FASTA was excluded. The influence of the PRED-FASTA exclusion did not appear in the 19 targets with GDT_TS evaluation.


1. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs *Nucleic Acids Res* **25**, 3389-3402.
2. Ogata,K. and Umeyama,H. (2000). An automatic homology modeling method consisting of database searches and simulated annealing *J Mol Graph Model* **18**, 258-272, 305-256.
3. Lo Conte,L., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2002). SCOP database in 2002: refinements accommodate structural genomics *Nucleic Acids Res* **30**, 264-267.


**FAMS** (serv) - 320 models for 64 3D targets

## Full automatic homology modeling server including the transformation of amino acid residues: SKE(Sophia Kai Ergon) FAMS


M. Iwadate[1], K. Kanou[1], G. Terashi[1], D. Takaya[1], M. Takeda-Shitaka[1] and H. Umeyama[1]

[1] - *Department of Biomolecular Desig*

*School of Pharmaceutical Sciences, Kitasato University*

iwadatem@pharm.kitasato-u.ac.jp


Selecting alignment

The aliment selection for constructing highly accurate protein models using homology modeling was executed for 8 kinds of methods, BLAST[1], PSI-BLAST, PSF-BLAST, RPS-BLAST, IMPALA, FASTA, Pfam and PRED-FASTA.

PSF-BLAST is PSI-BLAST whose profile sequence group of PSSM construction process is revised, and the selection criterion is E-value<=0.001 from template PDB sequence on PSI-BLAST search.

PRED-FASTA is unique and simple homologue detection program which 20 amino acid residues were transformed based on secondary structure and amino acid similarity. This program uses PSI-PRED[2] and FASTA[3].

In order to select the best in 8 kinds of alignment methods, the score-function that was constructed by model length, homology% and degree of secondary structure agreement between PSI-PRED and STRIDE was defined by the equation:

$$score = f(k_i, Hom, Len, SS)$$

$Len$ is residue length of model protein. $Hom$ indicate homology % value, the ratio between the number of match residues and $Len$. $SS$ is so called Q3 value, degree of secondary structure agreement between PSI-PRED and STRIDE. $k_i$ are coefficients. The subscript number "i" indicates the number of alignment method; 0 is PSI-BLAST, 1 is BLAST, 2 is RPS-BLAST, 3 is Family-BLAST, 4 is IMPALA, 5 is FASTA, 6 is PRED-FASTA, 7 is Pfam.

Selecting fragment

The fragment selection process of FAMS[4] was modified to select fragment of same protein family. Therefore selection criteria were RMSD of fitting, and degree of SCOP[5] ID agreement.

Energy Minimize and Molecular Dynamics

After homology modeling, both of Energy Minimization and Molecular Dynamics are applied.

Results and Discussion

In 29 available CASP6 target structures, models were evaluated with GDT_TS. Number of targets that maximum GDT_TS with 8 kinds of methods were more than 30 was 19. The 19 targets approximately correspond to "PDB-Blast hits" targets in CAFASP website, and 10 targets approximately correspond to no "PDB-Blast hits" targets. Therefore, 10 targets were high-difficulty targets in 29 targets. In these targets other criterion (visual inspection?) is required.

In 16 of 19 targets, the alignments that GTD_TS are more than 87% of maximum GDT_TS were selected with this score-function. High detection capability for GDT_TS with a simple score-function was indicated.


1. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402.
2. Jones,D.T. (1999). Protein secondary structure prediction based on position-specific scoring matrices *J Mol Biol/J Mol Biol* **292**, 195-202.
3. Pearson,W.R. and Lipman,D.J. (1988). Improved tools for biological sequence comparison *Proc Natl Acad Sci U S A* **85**, 2444-2448.
4. Ogata,K. and Umeyama,H. (2000). An automatic homology modeling method consisting of database searches and simulated annealing *J Mol Graph Model* **18**, 258-272, 305-256.
5. Lo Conte,L., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2002). SCOP database in 2002: refinements accommodate structural genomics *Nucleic Acids Res.* **30**, 264-267.


## **Fischer** - 189 models for 64 3D targets

### Beautifying 3D-SHOTGUN models


D. Fischer

*Bioinformatics Center of Excellence, University of Buffalo*

dfischer@bioinformatics.buffalo.edu


C-alpha, unrefined models were automatically generated for all targets using the 3D-SHOTGUN server. 3D-SHOTGUN assembles hybrid, C-alpha-only, unrefined models that can be an excellent starting point for refinement.

For CASP, we refined the 3D-SHOTGUN models using either Honig's nest or Keasar's beautify refinement programs. The resulting models are full-atom, physically valid models. Large tests from LiveBench indicate that the performance of these refinement programs in combination with 3D-SHOTGUN, is superior to that obtained by other refinement programs previously used. The MaxSub scores of the resulting models are on average almost identical to those obtained for the original 3D-SHOTGUN models. Thus, this procedure achieves two goals: accuracy and beauty. The first goal is achieved because 3D-SHOTGUN produces excellent unrefined models. The second goal is achieved because the refinement, without decreasing accuracy, produces physically valid, full-atom models. For verification purposes, the resulting full-atom models were assessed using the new MQAP-CONSENSUS method developed for MQAP-CAFASP (www.cs.bgu.ac.il/~dfischer/CAFASP4). The MQAP-CAFASP scores of our refined models were also compared to models produced by the CAFASP servers.

Results from the evaluation of the nearly 30 CASP6 targets whose structures have been released indicate that the beautified 3D-SHOTGUN models submitted to CASP6 are of relatively high quality: their total MaxSub score is higher than that obtained by the best CAFASP4 servers and meta-servers.

**Floudas** - 60 models for 12 3D targets

## ASTRO-FOLD: first principles tertiary structure prediction

C.A. Floudas, J.L. Klepeis, and S.R. McAllister
*Department of Chemical Engineering, Princeton University, Princeton, NJ*
floudas@titan.princeton.edu

ASTRO-FOLD is an integrated methodology for the first principles structure prediction of proteins based on an overall deterministic global optimization framework coupled with mixed-integer optimization. The novel four-stage approach combines the classical and new views of protein folding, while using free energy calculations and integer linear optimization to predict the location of helical segments and the topology of beta-sheet structures and disulfide bridges, respectively. Detailed atomistic-level energy modeling and the deterministic global optimization method, αBB, coupled with torsion angle dynamics, form the basis for the final tertiary structure prediction[1-4].

The first stage of the approach involves the identification of helical segments. This is accomplished through detailed atomistic-level energy modeling of overlapping subsequences of the overall protein sequence using the selected force field (e.g., ECEPP/3[5]). The amino acid sequence is first decomposed into subsequences of overlapping oligopeptides (e.g., pentapeptides, heptapeptides, nonapeptides). For instance, using heptapeptides, the folowing subsequences are generated: 1-7, 2-8, 3-9, . . . etc. For each subsequence, global optimization can be used to generate an ensemble of low energy conformations along with the global minimum energy conformation[6]. Rigorous free energies that include entropic, cavity formation, polarization and ionization contributions, and involve solution of the Poisson-Boltzmann equation, are calculated for a subset of conformations for each oligopeptide system. Finally, these free energy values are combined to determine helical propensities for each residue by calculating equilibrium occupational probabilities for each possible helical cluster[7].

The second stage focuses on the prediction of beta-sheet and disulfide bridge topology through the analysis of amino acid properties that are based on residue hydrophobicities. The approach, which borrows key concepts from a mathematical framework developed in the area of process synthesis of chemical systems[8], is based on the idea that beta-structure formation relies on a hydrophobic driving force. To model this force, it is necessary to predict contacts between hydrophobic residues. The first important component of the approach is the postulation of a beta-strand superstructure that encompasses all alternative beta-strand arrangements. A novel mathematical model is then formulated to provide the formation of ordered structural features, such as beta-sheets and disulfide bridge connectivity. The solution of this integer linear programming problem, with the objective being the maximization of the hydrophobic contact energy, provides a rank ordered list of preferred hydrophobic residue contacts, beta strand topologies and disulfide bridge connectivities[9].

The third stage involves the derivation of restraints based on helical and beta-sheet predictions in the form of dihedral angle and atomic distance restraints to enforce the predicted secondary and tertiary arrangements. In addition, restraints are developed by prediction of interhelical contacts for all alpha-helical proteins[10]. By maximizing the occurrence of highly probable pairwise interactions, a rank ordered list of helical topologies is produced using a detailed optimization model. Also, additional restraints can be determined for the intervening loop residues connecting helical and strand regions through novel application of free energy simulation[11-13]. More specifically, the identified loops are extended on each side to incorporate three additional amino acids of both secondary structure elements that the loop connects. Each set of three flanking amino acids are imposed to be in their respective secondary structure state (e.g., helix, beta-strand). Then, a series of free energy calculations are conducted using overlapping oligopeptides, similar to the free energy calculations in the helix prediction stage. The objective of these calculations is to produce improved bounds on the dihedral angle and backbone distances within the loop residues. However, due to the restrictive deadlines of the CASP6 competition, it became infeasible to apply these loop modeling efforts to the protein targets.

The fourth and final stage of the approach involves the prediction of the tertiary structure of the full protein sequence. The problem formulation, which relies on dihedral angle and atomic distance restraints introduced from the previous stages, as well as on detailed atomistic energy modeling, represents a nonconvex constrained global optimization problem. This problem is solved through the combination of a deterministic global optimization approach, the αBB method; a stochastic algorithm, conformational space annealing; and a preprocessing torsion angle dynamics step[1,4,13]. The resulting low energy ensemble is evaluated through a clustering analysis. A variant of a k-means algorithm predicts five clusters of conformers using protein C-alpha coordinates[14]. A distributed computing framework of each stage of the proposed approach has been developed, and our predictions in the CASP6 competition employ this parallel implementation.

1. Klepeis,J.L. & Floudas,C.A. (2003) Ab Initio Tertiary Structure Prediction of Proteins, *J. Global Optim.* **25**, 113-140.
2. Floudas, C.A. (2000) Deterministic Global Optimization: Theory, Algorithms and Applications, Kluwer Academic Publishers.
3. Klepeis,J.L., Schafroth,H.D., Westerberg,K.M. & Floudas,C.A. (2002) Deterministic global optimization and ab-initio approaches for the structure prediction of polypeptides, dynamics of protein folding, and protein-protein interactions, *Adv. Chem. Phys.* **120**, 265-457.
4. Klepeis,J.L. & Floudas,C.A. (2003) ASTRO-FOLD: A Combinatorial and Global Optimization Framework for Ab Initio Prediction of Three-Dimensional Structures of Proteins from the Amino Acid Sequence, *Biophys. J.* **85**, 2119-2146.
5. Nemethy,G., Gibson,K.D., Palmer,K.A., Yoon,C.N., Paterlini,G., Zagari,A., Rumsey,S. & Scheraga,H.A. (1992) Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm with applications to proline-containing peptides, *J. Phys. Chem.* **96**, 6472-6484.
6. Klepeis, J.L. & Floudas,C.A. (1999) Free energy calculations for peptides using deterministic global optimization, *J. Chem. Phys.* **110**, 7491-7512.
7. Klepeis,J.L. & Floudas,C.A. (2002) Ab-Initio Prediction of Helical Segments in Polypeptides, *J. Comp. Chem*, **23**, 1-22.
8. Floudas, C.A. (1995) Nonlinear and Mixed-Integer Optimization: Fundamentals and Applications, Oxford University Press.
9. Klepeis,J.L. & Floudas,C.A. (2003) Prediction of Beta-Sheet Topology and Disulfide Bridges in Polypeptides, *J. of Comp. Chem.* **24**, 191-208.
10. Mickus,B., Klepeis,J.L., McAllister,S.R., & Floudas,C.A. (2004). A Novel Approach for Alpha-Helical Topology Prediction. In preparation.
11. Klepeis,J.L. & Floudas,C.A. (2004) Analysis and Prediction of Loop Segments in Protein Structures, *Comput. Chem. Eng.*, in press.
12. Klepeis,J.L., Pieja,M.J. & Floudas,C.A. (2003) A New Class of Hybrid Global Optimization Algorithms for Peptide Structure Prediction: Integrated Hybrids, *Comput. Phys. Comm.* **151**, 121-140.
13. Klepeis,J.L., Pieja,M.J. & Floudas,C.A. (2003) Hybrid Global Optimization Algorithms for Protein Structure Prediction: Alternating Hybrids, *Biophys. J.* **84**, 869-882.
14. Moennigmann,M., McAllister,S.R. & Floudas,C.A. Unpublished.


**FORTE1** (serv) - 320 models for 64 3D targets

## FORTE1: a simple profile-profile comparison method applied to fold recognition


K. Tomii

*Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, 2-43 Aomi, Koto-ku, Tokyo, Japan*

k-tomii@aist.go.jp


In the CASP5 experiment, we proposed a simple fold recognition technique and built its automated server, FORTE1, based on a profile-profile comparison method. The FORTE1 server had been given a comparatively higher rank by virtue of some evaluation results[1–2], and was included in Pcons-5 at LiveBench. The server[3] is publicly available for academic use now. This approach has also been applied to protein structure prediction of the CASP6 targets, but its profile database has been improved, as explained below.

The FORTE1 system uses position-specific score matrices (PSSMs) of both the query and templates as profiles. It identifies proper templates and produces profile-profile alignments of a target and templates. To calculate PSSMs of both the query and templates, PSI-BLAST[4] iterations are performed a maximum of 20 times with the NCBI non-redundant database. The amino acid sequences of templates are derived from the ASTRAL[5] 40% identity list and selected PDB[6] entries that are not registered in the SCOP[7] database. Furthermore, the full-length sequences, which are divided into structural domains in SCOP, are also prepared because we believe our prediction results were not reflected appropriately in some cases in CAFASP3. Our server recognizes correct domains separately (typically in the case of T0185).

The standard dynamic programming algorithm is used with gap penalties that are optimized by our preliminary study, explained below, to align two PSSMs. The dynamic programming algorithm requires a matrix containing similarity scores for the pairs of positions in the PSSMs that are to be compared. The similarity score for each pair of PSSM columns is defined as their correlation coefficient. We use the global alignment algorithm with no penalty for the terminal gaps to obtain an optimal sequence-structure alignment. The statistical significance of each alignment score is estimated by calculating the Z-scores with a simple log-length correction. Candidates of sequence-structure alignments were sorted by their Z-scores. We submitted prediction results in the AL format.

We employ the Pearson's correlation coefficient to measure the similarity between two profile columns, as stated previously, because we have found that, compared with the dot product, the correlation coefficient offers an advantage. It showed higher sensitivity of fold recognition at the SCOP family, superfamily, and fold level, when we performed our preliminary study with 948 single domain proteins selected by PDB-REPRDB[8]. Thereby, any pair of proteins has less than 30% sequence identity. A similar tendency was noted in a recent paper[9].


1. Wallner,B., Fang,H., & Elofsson,A. (2003). Automatic consensus-based fold recognition using Pcons, ProQ, and Pmodeller. *Proteins **53***, 534-541.
2. http://supfam.mrc-lmb.cam.ac.uk/CAFASP/
3. Tomii,K. & Akiyama,Y. (2004). FORTE: a profile-profile comparison tool for protein fold recognition. *Bioinformatics **20***, 594-595.
4. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res. **25***, 3389-3402.
5. Chandonia,J.M., Hon,G., Walker,N.S., Lo Conte,L., Koehl,P., Levitt,M. & Brenner,S.E. (2004). The ASTRAL compendium in 2004. *Nucleic Acids Res. **32***, D189-D192.
6. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. & Bourne,P.E. (2000). The Protein Data Bank. *Nucleic Acids Res. **28***, 235-242.
7. Andreeva,A., Howorth,D., Brenner,S.E., Hubbard,T.J.P., Chothia,C. & Murzin,A.G. (2004). SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res. **32***, D226-D229.
8. Noguchi,T. & Akiyama,Y. (2003) PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB) in 2003. *Nucleic Acids Res. **31***, 492-493.
9. Wang,G. & Dunbrack,R.L. Jr. (2004) Scoring profile-to-profile sequence alignments. *Protein Sci. **13***, 1612-1626.


**FORTE1T** (serv) - 320 models for 64 3D targets

## FORTE1T: profile-profile comparison with improved profiles for fold recognition


K. Tomii

*Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, 2-43 Aomi, Koto-ku, Tokyo, Japan*

k-tomii@aist.go.jp


MOTIVATION: We have constructed a new server, FORTE1T, to elucidate effects of quality of profiles for alignment accuracy as well as sensitivity and selectivity of fold recognition. This system uses a process that produces a multiple alignment as a seed, thereby improving profile quality before profile construction by PSI-BLAST[1] iterations.

PROFILE CONSTRUCTION FOR FORTE1T: Amino acid sequences of templates are derived from the ASTRAL[2] 40% identity list and selected PDB[3] entries which are not registered in the SCOP[4] database. Furthermore, full-length sequences that are divided into structural domains in SCOP are also prepared. Those are template library sequences. Moderately related sequences for each sequence are gathered by PSI-BLAST from the NCBI non-redundant database. A multiple alignment is constructed for each template using T-Coffee[5] with those related sequences and the template sequence. Multiple alignment is used as the seed alignment for profile construction by PSI-BLAST iterations with the NCBI non-redundant database. This process to improve profile quality is applied for both the query and templates.

BENCHMARK RESULT: To evaluate the ability of FORTE1T for fold recognition, FORTE1T was also included partly in LiveBench-9. For most cases, both alignment accuracy and sensitivity of fold recognition were improved over those of FORTE1[6].

IN CASP6: The FORTE1T system also uses position-specific score matrices (PSSMs) of both the query and templates as profiles to predict the structure of the query sequence. Except for the profile quality, the procedures to obtain candidates of sequence-structure alignments are identical to those of FORTE1. We submitted prediction results in the AL format.


1. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res. **25***, 3389-3402.

2. Chandonia,J.M., Hon,G., Walker,N.S., Lo Conte,L., Koehl,P., Levitt,M. & Brenner,S.E. (2004). The ASTRAL compendium in 2004. *Nucleic Acids Res.* ***32***, D189-D192.
3. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. & Bourne,P.E. (2000). The Protein Data Bank. *Nucleic Acids Res.* ***28***, 235-242.
4. Andreeva,A., Howorth,D., Brenner,S.E., Hubbard,T.J.P., Chothia,C. & Murzin,A.G. (2004). SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.* ***32***, D226-D229.
5. Notredame,C., Higgins,D.G. & Heringa,J. (2000). T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* ***302***, 205-217.
6. Tomii,K. & Akiyama,Y. (2004). FORTE: a profile-profile comparison tool for protein fold recognition. *Bioinformatics* ***20***, 594-595.

## **FORTE2** (serv) - 320 models for 64 3D targets

### FORTE2: automated fold recognition server with enhanced profile library

K. Tomii
*Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, 2-43 Aomi, Koto-ku, Tokyo, Japan*
k-tomii@aist.go.jp

MOTIVATION: To elucidate effects including very distantly related sequences into profiles for alignment accuracy, as well as sensitivity and selectivity of fold recognition, we have constructed our new server: FORTE2 (FORTE is an abbreviation for "FOld Recognition TEchnique"). Its system uses the same protocol as FORTE1[1]. It has enriched profiles by incorporating highly diverged sequences detected by FORTE1 into the sets of sequences that are gathered by PSI-BLAST[2].

PROFILE CONSTRUCTION FOR FORTE2: First, protein domain sequences were derived from a 40% identity list of SCOP[3] 1.63. Their profiles were constructed using the FORTE1 procedure. Those sequences and profiles were prepared as a representative data set. Through an all-against-all search of this data set by FORTE1, we identified the true positive pairs of proteins with Z-score, ranging from 4 to 10. In this case, we determined true positive pairs as those proteins that are assigned the same fold in the SCOP classification. We constructed new profiles using alignments of those pairs for FORTE2. Those alignments, provided by FORTE1, were used as seed alignments for profile construction by PSI-BLAST iterations with the NCBI non-redundant database.

BENCHMARKS: We performed an all-against-all search using the representative data set to evaluate and compare the ability of identifying proteins with the same fold by FORTE1 and FORTE2. In this test, we regarded relationships with higher Z-scores than the first false positive as true hits. FORTE1 and FORTE2 were also included in LiveBench.

BECHMARK RESULTS[4]: We found that subtle effects of incorporating highly divergent sequences detected by FORTE1 into the sets of sequences that had been gathered by PSI-BLAST in profile construction. We found that FORTE2 can detect relationships between proteins that are different from those detected by FORTE1 through all-against-all search, but most true hits are common to both methods. According to LiveBench-8 results, FORTE2 showed the additional advantage of remote homology or analogy detection, but with slightly worsened alignment accuracy in some cases.

IN CASP6: The FORTE2 system also uses position-specific score matrices (PSSMs) of both the query and templates to predict the structure of the query sequence, as FORTE1 does. The enhanced profile library was updated. Procedures to obtain an optimal sequence-structure alignment and estimate its statistical significance are the same as those of FORTE1. Candidates of the sequence-structure alignments were sorted by their Z-scores. Subsequently, we submitted prediction results in the AL format.

1. Tomii,K. & Akiyama,Y. (2004). FORTE: a profile-profile comparison tool for protein fold recognition. *Bioinformatics* ***20***, 594-595.
2. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* ***25***, 3389-3402.
3. Murzin,A. Brenner,S.E., Hubbard,T. & Chothia,C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* ***247***, 536-540.
4. Tomii,K. (2004). Effects of using highly diverged sequences in profile-profile comparisons. The 1st Pacific-Rim International Conference on Protein Science P17/18-217.

**FRCC** - 56 models for 51 3D / 2 DP targets

## EVM (Expert vs Machine) strategy for fold recognition

I.Y. Torshin
*Private informatics consulting*
tiy135@yahoo.com

There has seen an enormous increase in the number of the individual computerized methodologies for protein structure prediction. Entirely automated methods are most appropriate when applied to the tasks of the comparative modeling rather than to fold recognition or "*ab initio*"-styled algorithms. Experimentation with different fold recognition programs (such as 3D-PSSM[1], FFAS[2], FoldRec-CC[3] and others) shows that at low sequence similarities structures of very different architectures are often selected among the 10,20,30... "top-ranking" and the correct template is often on the list. However, the sequence-based E-value does not in any way discriminate between the correct and the incorrect templates. This is the stage of prediction when the human's expertise, of whatever level, should of necessity come into view.

On one hand, it is the author's conviction that a method of protein structure prediction applicable to wide range of sequences (with any identities to known sequences) should be based and consequently developed on the fundamental natural laws. So far, there were quite a few attempts to develop methods (for example, "Rosetta" [4]) in this extraordinarily important direction of research and most of the methods appear to be content with various statistical models. Statistical models are, in general, only a surrogate for, or, at best, a semi-product of the actual scientific understanding[5]. As shows the centuries-long history of physics, the fundamental laws can always be implemented as reliable computational procedures. On the other hand, there is a possibility that the *structural proteomics* (sometimes weirdly called as "structural genomics", to the considerable annoyance of geneticists and microbiologists) will yield sufficient templates to model any protein structure. However, this is definitely not the case at present.

Thus, a computational procedure that has been cleansed from the arbitrary assumptions and that is supplemented by a human expertise can be a practical solution to the fold recognition problem. It becomes increasingly clear that secondary structure is a fundamental characteristics of proteins[6-8] and its important role in the folding also becomes apparent[7]. Therefore, moving the accents of the similarity searches from "primary" to "secondary" structure similarity allows a jump through the problem of low sequence identity (at least in theory). Such a procedure should not rely on the sequence identity (in any of its multiple forms and definitions) as the sole criterion for template selection.

The algorithm we applied to CASP6 targets was elaborated largely on the base of the above considerations. At the first stage, the secondary structure was predicted (Psipred[9]) and hashed against a non-redundant database of about 5,000 templates. The best-matches (300-500) were re-ranked according to the compactness and domain isolation. Domain definitions were from GTDD[3]. The 3D models were prepared using the final list of matches (<20 models) and CLUSTAL W alignments. Models were annotated with the function and domain predictions, SCOP and GTDD domains. The final list was carefully analyzed to select the most appropriate templates for the target sequence, appropriateness of each potential template was determined through the human expertise including "LSH-calculus" developed by the author. Only targets with such "appropriate templates" were submitted to the CASP-6 experiment.

It is quite obvious that the accuracy of the entire procedure depends on the accuracy of the secondary structure prediction (at present 80%, and this is an arguable average). Of course, the procedure incorporates correction algorithms accounting for possible errors in the secondary structure prediction. However, experimenting with secondary structures of different accuracies shows that the accuracy should be not less than 80%-85% to produce an accurate 3D model. Another problem that arises especially at lower accuracies of the secondary structure prediction is the requirement of the variety of the secondary structure elements along the amino acid sequence. For example, procedure would not work well on the structures with long loops and very few and short strands, such as found, for example, in kringle and EGF-like domains. Comparisons with predictions by other methods (primarily, BLAST, 3DPSSM and FFAS) show strong agreement for a number of CASP6 targets, although at present (Oct 2004) the actual efficiency of the method in its current form is arguable as the targets' info is not yet available.

1. Kelley,L.A., MacCallum,R.M., Sternberg,M.J (2000). Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J Mol Biol.* **299**, 499-520.
2. Rychlewski,L., Jaroszewski,L., Li,W., Godzik,A. (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci* **9,** 232-241.
3. Torshin,I.Y. (2002). Net charge center for protein fold recognition, Proc. 5th CASP meeting, Azilomar, CA, Dec 2002, p. A-163.
4. Bradley,P., Chivian,D., Meiler,J., Misura,K., Rohl,C., Schief,W., Wedemeyer,W., Schueler-Furman,O., Murphy,P., Schonbrun,J., Strauss,C.,

Baker,D. (2003) Rosetta predictions in CASP5: successes, failures, and prospects for complete automation. *Proteins* **6**, 457-468.
5. D. Bohm (1987) Causality and Chance in Modern Physics (with preface by L. de Broigle), University of Pennsylvania Press.
6. Torshin,I.Y., Harrison,R. (2003) Protein folding: search for basic physical models. *ScientificWorldJournal* **3,** 623-635.
7. Torshin,I.Y. Computed energetics of macromolecules: proteins and RNA. (2004) 227th meeting of American Chemical Society, Anaheim, CA, USA, A345
8. Torshin,I.Y. Computed energetics of macromolecules: the roles of the subsequent levels of structure, part II (2004). In: "Progress in Bioinformatics", in print.
9. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol. **292***, 195-202.

**FUGMOD_SERVER** - 320 models for 64 3D targets

## FUGMOD: an automatic comparative modelling server

Ricardo Núñez Miguel, Tom L. Blundell and Kenji Mizuguchi
*Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, Cambridge CB2 1GA*
kenji@cryst.bioc.cam.ac.uk

FUGMOD is a server that runs the Automatic COMParative MODelling program ACOMPMOD. The program consists of several components that run external fold recognition/comparative modelling programs and efficiently analyse and combine their output. This makes ACOMPMOD a powerful tool for fully automated comparative modelling. The only needed input is the amino acid sequence of the protein of interest.

ACOMPMOD utilises the homology recognition program FUGUE[1]. It takes the alignments produced by FUGUE and then runs the program ALIMOD, which modifies the alignments in order to place the deletions in the most appropriate location within the target sequence by calculating the spatial distances between the two residues of the template that are aligned with the residues at the borders of the deletion in the target sequence. In the next step, ACOMPMOD runs MODELLER[2,3] to build atomic coordinates. In this step nine models are solicited. ACOMPMOD selects the best model using the program MODELLIST. This program utilises the energy and violations of every model in order to obtain values that are considered in the selection of the best model. Once the best model is selected, ACOMPMOD runs the program JOY[4] to annotate protein sequence alignments with three-dimensional (3D) structural features. The JOY output will help in the validation of the model.

ACOMPMOD produces full-atom models with all residues, including loops, excluding, sometimes, the N- and/or C-termini if they do not have templates to be aligned with.

Different options are available when running ACOMPMOD. The program can accept only an amino acid sequence, run FUGUE and produce models for 1) the top ten FUGUE hits or 2) only for the single highest scoring one. It can also accept an amino acid sequence and a HOMSTRAD[5-7] family to produce a model using that family as template. Furthermore, it can accept an amino acid sequence, a HOMSTRAD family and an alignment to produce a model based on that family and the user-supplied alignment. In all cases the user can choose whether to use the ALIMOD for modifying the alignment or not.

1. Shi,J., Blundell,T.L. & Mizuguchi,K. (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol.* **310**, 243-257
2. Sali,A. & Blundell,T.L. (1993) Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J. Mol. Biol.* **234**, 779-815.
3. Fiser,A., Do,R.K. & Sali,A. (2000) Modeling of loops in protein structures. *Protein Sci.* **9,** 1753-1773.
4. Mizuguchi,K., Deane,C.M., Blundell,T.L., Johnson,M.S. & Overington,J.P. (1998) JOY: protein sequence-structure representation and analysis. *Bioinformatics* **14**, 617-623.
5. Mizuguchi,K., Deane,C.M. & Blundell,T.L. (1998) HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.* **7**, 2469-2471.
6. deBakker,P.I.W., Bateman,A., Burke,D.F., Miguel,R.N., Mizuguchi,K., Shi,J., Shirai,H. & Blundell,T.L. (2001) HOMSTRAD: adding sequence information to structure-based alignments of homologous protein families. *Bioinformatics* **17**, 748-749.
7. Stebbings,A.L. & Mizuguchi,K. (2004) HOMSTRAD: Recent developments of the Homologous Protein Structure Alignment Database. *Nucleic Acids Res.* **32**, D203-D207.

## **FUGUE_SERVER** - 320 models for 64 3D targets

### FUGUE – recent enhancements to sequence-structure homology recognition

Jiye Shi, Tom L. Blundell and Kenji Mizuguchi
*Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, Cambridge CB2 1GA*
kenji@cryst.bioc.cam.ac.uk

The key elements of the homology recognition software FUGUE[1], when first tested in CASP3[2], were environment-specific substitution tables (ESSTs), structure-dependent gap penalties, automated alignment selection and the use of the HOMSTRAD[3-4] database, a curated collection of protein structural families. A position-specific score matrix (PSSM) was derived using the ESSTs from the structure-based alignment of each family in HOMSTRAD. Homologues of the query sequence were collected and the resulting sequence profile was compared against the structure-based PSSMs. A new feature was introduced and tested in CASP4 and CASP5, to enrich the structure-based PSSMs with information derived from homologous sequences. This enhancement improved the performance dramatically[5] and it has now become the default feature. Therefore, FUGUE not only utilizes structural information (in the form of ESSTs), as in many fold recognition methods, but also incorporates the elements of sequence-based structural profile enrichment and profile-profile alignment techniques[6].

Although the program has proved successful in other benchmark exercises and continues to uncover novel homologies[7-9], our own recent benchmark results suggested that the structural and sequence information may not be optimally combined in the enriched profiles. This was because the enriched PSSMs did not always produce better alignments than the original PSSMs (with no added homologous sequences). The homologous sequences were added by PSI-BLAST to the structure-based alignments and the quality of the PSI-BLAST alignments appeared to influence the efficiency of structural profile enrichment. A new algorithm was recently introduced to filter the PSI-BLAST output and it has improved the performance of FUGUE with enriched PSSMs in our benchmark exercises. This new option is being further tested in CASP6.

1. Shi,J., Blundell,T.L. & Mizuguchi,K. (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol.* **310**, 243-257.
2. Burke,D.F. et al. (2000) An iterative structure-assisted approach to sequence alignment and comparative modelling. Proteins Suppl **3,** 55-60.
3. Mizuguchi,K., Deane,C.M. & Blundell,T.L. (1998) HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.* **7**, 2469-2471.
4. Stebbings,A.L. & Mizuguchi,K. (2004) HOMSTRAD: Recent developments of the Homologous Protein Structure Alignment Database. *Nucleic Acids Res.* **32**, D203-D207.
5. Williams, M.G. et al. (2001) Sequence-structure homology recognition by iterative alignment refinement and comparative modelling. Proteins Suppl **5**, 92-97.
6. Mizuguchi K. (2004) Fold recognition for drug discovery. Drug Discovery Today: Targets **3**, 18-23.
7. Witty, M. Sanz, C., Shah, A., Grossmann, J.G., Mizuguchi, K., Perham, R.N., Luisi, B. (2002) Structure of the periplasmic domain of Pseudomonas aeruginosa TolA: evidence for an evolutionary relationship with the TonB transporter protein. *EMBO J.* **21**, 4207-4218.
8. Nishi J, Sheikh J, Mizuguchi K, Luisi B, Burland V, Boutin A, Rose DJ, Blattner FR, Nataro JP. (2003) The export of coat protein from

enteroaggregative Escherichia coli by a specific ATP-binding cassette transporter system. *J Biol Chem* **278**, 45680-9.
9. Shirai H, Mizuguchi K. (2003) Prediction of the structure and fucntion of AstA and AstB, the first two enzymes of the arginine succinyltransferase pathway of arginine catabolism. *FEBS Lett* **555**, 505-510.

## **GeneSilico-Group** - 192 models for 64 3D targets

### Dr. Frankenstein's toolbox

M.J. Gajda, J. Kosinski, I.A. Cymerman, M.A. Kurowski, M. Pawlowski, M. Boniecki, A. Obarska, G. Papaj, P. Sroczynska, K. Tkaczuk, P. Sonta, A. Augustyn, J.M. Bujnicki and M. Feder
*International Institute of Molecular and Cell Biology, Trojdena 4, 02-109 Warsaw, Poland*
marcin@genesilico.pl

During the previous CASP5 experiment, our group (GeneSilico/517) applied a multi-step protocol called "FRankenstein Monster's approach" to predict protein structures of all targets regardless of their potential classification[1]. This strategy used multiple models from various fold-recognition (FR)-servers to construct a hybrid model and tweak it locally to obtain an optimal target-template alignment. According to the official assessment we were able to build very accurate models in the comparative modelling (CM)[2] and easy fold recognition (FR/H)[3] categories targets but our performance was poorer for hard fold recognition (FR/A) targets and we failed to predict correct structures for all but one target in the new fold (NF) category. Another relevant difficulty we met was reconstructing large insertions in the target without any counterpart in the potential templates.

Several limitations of "FRankenstein's Monster" approach were avoided in CASP6. Previously "alignment shifting" and subsequent evaluation was done manually. In CASP6 edition we automatized this steps what allowed us to probe much higher number of possible alignments and only the best scoring ones underwent manual and knowledge-based inspection.

The most severe limitation of the previous implementation of the "Frankenstein's Monster" approach in CASP5 was a lack of a reasonable method to predict conformation of long loops, large insertions, terminal extensions and domains with no template identified. In CASP6, we applied Rosetta[4] to model *de novo* fragments of models for which the conformation could not be confidently inferred from the templates. It was especially useful when only the protein core could be modeled by comparative modeling, while loops and peripheral elements had to be modeled *de novo*. In the cases we identified as potential new folds, we used Rosetta in a fully *de novo* mode. However, if we could identify any structural similarity between Rosetta models and some of the potential templates, the templates were used preferentially.

Summarizing, in CASP6 we used the FRankenstein.s approach in similar manner as in CASP5, but in a more automated and extended fashion. Most analyses were carried out using components of a fully automated package of programs and libraries (Dr. Frankenstein's Toolbox), which in the near future will be made available to the scientific community as a WWW server.

1. Kosinski,J., Cymerman,I.A., Feder,M., Kurowski,M.A., Sasin,J.M. & Bujnicki,J.M. (2003) A "FRankenstein's monster" approach to comparative modelling: merging the finest fragments of Fold-Recognition models and iterative model refinement aided by 3D structure evaluation. *Proteins* **53(S6)**:369-379.
2. Tramontano,A. & Morea,V. (2003). Assessment of homology-based predictions in CASP5. *Proteins* **53(S6)**:352-368.
3. Kinch,L.N., Wrabl,J.O., Krishna, S.S., Majumdar, I., Sadreyev, R.I., Qi,Y., Pei,J., Cheng,H. & Grishin,N.V. (2003) CASP5 assessment of fold recognition target predictions. *Proteins* **53(S6)**:395-409.
4. Rohl, C. A., Strauss, C. E., Misura, K. M. & Baker, D. (2004). Protein structure prediction using Rosetta *Methods Enzymol* **383**, 66-93.

## **Ginalski** - 150 models for 64 3D targets

### Modeling of CASP6 target proteins with 3D-Jury, Meta-BASIC and ROSETTA

K. Ginalski
*Department of Biochemistry, University of Texas, Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, Texas 75390-9038, USA*
kginal@chop.swmed.edu

For the sixth round of Critical Assessment of Techniques for Protein Structure Prediction (CASP6), 76 target proteins were modeled based on the results of 3D-Jury[1], a consensus method of fold recognition servers, Meta-BASIC[2], a novel meta profile alignment method, and an ab-initio ROSETTA program[3].

The critical steps in comparative and fold recognition modeling: selection of template(s) and generation of sequence-to-structure alignment, were guided mainly by the results of secondary structure prediction and tertiary fold recognition carried out using the Meta Server[4]. Initially, related proteins with known structures were identified from the transitive PSI-Blast[5] searches performed against the NCBI non-redundant protein sequence database until profile convergence and from the consensus of the Meta Server results. For difficult targets, template/fold identification was based on the results of the 3D-Jury method for the query sequence and a few homologues as well as the transitive Meta-BASIC searches performed against PfamA, PfamB and PDB. In addition, fold selection was also guided by the consensus of exhaustive fold recognition searches with Meta-BASIC for sets of homologues detected with PSI-Blast (E-value <10) and by the similarity of ROSETTA models to known structures detected with MAMMOTH program[6]. Structural determinants of the selected folds were then analyzed: representative structures of a given fold and the corresponding structural alignments were inspected for both conservation and variability of the structural elements. Conservations of specific residues and contacts responsible for maintaining tertiary structure and critical for substrate binding and/or catalysis were also established. Additionally, PCMA program[7] was used to generate multiple sequence alignments for target and template families. Sequence-to-structure alignments were built using the consensus alignment approach and 3D assessment[8] within the context of the structural and sequential constraints identified above. Importantly, in this procedure several alignment variants for the most questionable regions were derived manually, guided mainly by secondary structure predictions and conservation of structurally important residues in the PCMA family profiles. Final models of target proteins were built with the MODELLER program[9] using more than one template structure where possible. In many cases loops and structurally variable regions were built manually or taken from other distantly related structures. No energy minimization procedures were employed.

For the remaining targets initial models were generated with the ROSETTA program for both the query protein and a few homologues as well. Resulting models were inspected manually and final selection was based mainly on consistency and compactness of the predicted structure. In several cases final models were built manually from the most common fragments extracted from ROSETTA models, taking into account predicted secondary structure, hydrophobic profile of the family and the location of absolutely conserved and presumably catalytic residues.


1. Ginalski,K., Elofsson,A., Fischer,D. & Rychlewski,L. (2003). 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics* **19**, 1015-1018.
2. Ginalski,K., von Grotthuss,M., Grishin,N.V. & Rychlewski,L. (2004). Detecting distant homology with Meta-BASIC. *Nucleic Acids Res.* **32**.
3. Bonneau,R., Tsai,J., Ruczinski,I., Chivian,D., Rohl,C., Strauss,C.E. & Baker,D. (2001). Rosetta in CASP4: progress in ab initio protein structure prediction. *Proteins* **Suppl 5**, 119-126.
4. Bujnicki,J.M., Elofsson,A., Fischer,D. & Rychlewski,L. (2001). Structure prediction meta server. *Bioinformatics* **17**, 750-751.
5. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
6. Ortiz,A.R., Strauss,C.E. & Olmea,O. (2002). MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.* **11**, 2606-2621.
7. Pei,J., Sadreyev,R. & Grishin,N.V. (2003). PCMA: Fast and Accurate Multiple Sequence Alignment Based on Profile Consistency. *Bioinformatics* **19**, 427-428.
8. Ginalski,K. & Rychlewski,L. (2003). Protein structure prediction of CASP5 comparative modeling and fold recognition targets using consensus alignment approach and 3D assessment. *Proteins* **53 Suppl 6**, 410-417.
9. Sali,A. & Blundell,T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779-815.


## **Glo4** - 2 models for 1 3D target

### Global search on a detailed energy surface


M.J. Dudek[1]

[1] – *Recently unemployed.*

mdudek@nethere.com


Structure prediction was attempted by global energy minimization of a detailed rigid-geometry protein energy function for the smallest CASP6 target T0215 consisting of 53 residues. Calculations were carried out on a 1.7 GHz Pentium IV processor. My two submitted structures correspond to steps 3 and 4 of a short trajectory of local minima consisting of 4 steps. Full global energy minimization was not possible due to limited computational resources.

The energy surface[2-4], currently unnamed, is based on a nonstandard collection of detailed functional forms. These include a distributed atomic multipole representation of the electrostatic component, a buf14-7 representation of the repulsion +dispersion component, a 2-dimensional fourier series representation of the intrinsic torsional component, and a hydration shell model representation of the hydrophobic contribution to hydration free energy. The remainder of hydration free energy is obtained as the energetic effect of a continuous dielectric medium calculated using a boundary element solution to the Poisson equation. Parameters were fit to small molecule data, including crystal structures and hydration free energies, obtained from both experimental measurements and molecular orbital calculations.

Initial validation of this energy surface[4] was through applications of global energy minimization to 7-residue surface loop segments of protein crystal structures. For 9 of 10 predictions, the native backbone conformation was identified correctly. The energy surface has continued to perform well in surface loop structure prediction of 9-residue segments, and in *ab initio* peptide structure prediction of omega-conotoxin family members ranging in size from 24 to 31 residues.

Global energy minimization is accomplished by generating a sequence, alternatively referred to as a trajectory, of local minima. Each step of this trajectory consists of generation of a large collection of starting conformations, intended to cover uniformly some subspace of possible deformations, followed by fast screening and local energy minimization. The local minimum conformation having the lowest energy is retained as the next step of the trajectory and the starting point for the next deformation.

The global search program, currently named GLO4, was originally developed for applications to protein surface loop segments ranging in length from 5-15 residues.[1] Feedback from global energy minimization of protein surface loops has motivated and guided development of the present better-performing energy surface. The GLO4 program has since been extended to enable applications of our detailed energy surface to structure prediction of peptides and small proteins. Recent extensions also enable applications to homology model building and ligand binding. It is hoped that feedback from small protein structure prediction can guide further improvement to the energy surface, although computational limitations remain a barrier.


1. Dudek,M.J. & Scheraga,H.A. (1990) Protein Structure Prediction Using a Combination of Sequence Homology and Global Energy Minimization I. Global Energy Minimization of Surface Loops. *J. Comput. Chem.* **11**, 121-151.
2. Dudek,M.J. & Ponder,J.W. (1995) Accurate Modeling of the Intramolecular Electrostatic Energy of Proteins. *J. Comput. Chem*. **16**, 791-816.
3. Dudek,M.J. & Hagler,A.T. (unpublished) The Impact of Atomic Dipoles and Quadrupoles on Calculated Crystal Structures and Sublimation Energies of Model Amide Compounds.
4. Dudek,M.J., Ramnarayan,K. & Ponder,J.W. (1998) Protein Structure Prediction Using a Combination of Sequence Homology and Global Energy Minimization II. Energy Functions. *J. Comput. Chem*. **19**, 548-573.
5. Dudek,M.J. & Ramnarayan,K. (2001) Application of a Detailed Energy Surface to Homology Modeling of the omega-Conotoxin Family. *Proceedings of the Seventeenth American Peptide Symposium* 428-429.


## **Hamilton-Huber-Torda** (serv) - 61 models for 61 RR targets

### Protein contact prediction using patterns of correlation


N.A. Hamilton[1,2], K. Burrage[1], M.A. Ragan[2], A.E. Torda[3] and T. Huber[1]

*1– Advanced Computational Modelling Centre, The University of Queensland,*
*2– Institute for Molecular Bioscience, The University of Queensland,*
*3 – Zentrum für Bioinformatik, Universität Hamburg*
n.hamilton@imb.uq.edu.au


Protein contact prediction provides a complementary approach to the information provided by force field and sequence alignment based methods for protein fold prediction. While the predictive accuracy is far from perfect it can provide valuable complementary information that can be used, for instance, to rank models created by other methods. In the following we describe a new method for contact prediction by training a Neural Network to classify patterns of contact. The main inputs to the neural network are a set of 25 measures of correlated mutation between all pairs of residues in two “windows” centered on the residues of interest. The individual pairwise correlations are a relatively weak predictor of contact, but by training the network on windows of correlation the accuracy of prediction is significantly improved.

Method

Psipred[4] version 2.3 software is used to generate a prediction for the secondary structure as well as giving a pair-wise multiple sequence alignment for the

proteins sequence. For each pair of residues in the protein sequence we generate a pattern of inputs for a neural network as follows.

*Pairwise correlations.* The multiple sequence alignment is used to calculate the (mutational) correlation between two columns of the multiple sequence alignment. The correlations are calculated as in Göbel et al.[1], with the minor modification that the Blosum62 matrix rather than that of McLachlan is used to score the residue interchanges. Windows of length 5 of consecutive columns are found. For each pair of non-overlapping windows the 25 correlations between columns of the first window with columns of the second are used as inputs to the neural network. The aim is to predict whether the middle residue of the first window is in contact with the middle residue of the second.

*Residue classes.* Residues may be classified as non-polar, polar, acidic, or basic. For a pair of residues there are ten possible pair cases. Thus we have ten binary inputs, exactly one of which is set to one to encode the residue type of the pair we are attempting to predict on.
*Predicted secondary structure.* For a given residue, its predicted secondary structure type is encoded as three binary inputs, being either helix, sheet or neither. For a given residue pair that we are attempting to predict with, the predicted secondary structure is input for the two residues as well as the two residues that are adjacent to them.

*Affinity score.* A given residue pair is assigned an affinity score based on the type of each of the amino acids. This expresses the fraction of times residue pairs of a given type are in contact in a training set of 50 proteins.

*Length of input sequence and residue separation.* The length of the sequence and the sequence separation, each divided by 1000, are input for the pair we are predicting with.

Network Architecture and Training
The predictor neural network is a standard feed-forward network, with 56 inputs, ten hidden units, and a single output. The expected output is 1 for contacts and 0 for non-contacts.

Proteins were randomly chosen from a representative set of proteins of the Protein Data Bank. The network was trained, validated and tested on disjoint sets of 100, 50 and 1033 proteins using back propagation with a momentum term with the Stuttgart Neural Network Simulator[5].

Testing the Trained Network

The trained network was tested on a set of 1033 proteins of known structure. An average predictive accuracy of 21.7% was obtained taking the best L/2 predictions for each protein, where L is the sequence length. Taking the best L/10 predictions gives an average accuracy of 30.7%. An automated prediction server can be found at

http://foo.maths.uq.edu.au/~nick/Protein/contact.html


1. Göbel,U., Sander,C., Scheider,R., Valencia,A. (1994) Correlated mutations and residue contacts in proteins. *Proteins* **18**, 309-317.
2. Fariselli,P., Olmea,O., Valencia,A., Casadio,R. (2001) Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations. *Proteins* ***Suppl* 5,** 157-162.
3. Hamilton,N., Burrage,K., Ragan,M., Huber,T. (2004) Protein contact prediction using patterns of correlation, *Proteins* ***56***, 679-684.
4. McGuffin,L.J., Bryson,K., Jones,D.T. (2000) The PSIPRED protein structure prediction server. *Bioinformatics* **16**, 404-405.
5. Zell,A., et al. (1998) Stuttgart neural network simulator user manual version 4.2. University of Stuttgart.


**HHpred.2** (serv) - 310 models for 62 3D targets
**HHpred.3** (serv) - 309 models for 62 3D targets

## Homology detection and 3D structure prediction by HMM-HMM comparison


J. Söding
*Dept for Protein Evolution,*
*Max-Planck-Institute for Developmental Biology, Tübingen, Germany*
johannes.soeding@tuebingen.mpg.de


The HHpred web server allows users to search for distant homologs of their query sequence in several databases like Pfam, SMART, or SCOP. It returns a list of best matches together with the query-template alignments in an easily readable format. We try to maximize flexibility for interactive use, providing the possibility to check the automatically generated alignment for errors or to search with a user-generated alignment. The server can be accessed at (http://protevo.eb.tuebingen.mpg.de/toolkit/index.php).

For participation in CASP6/CAFASP4, we developed a fully automated version of HHpred, called HHpred.2. It generates unrefined pdb-formatted structure models by using the query-template alignments to directly map the coordinates of the best templates to the query residues A multiple alignment is built from the query sequence using up to 8 rounds of PSI-BLAST with E-value threshold 1E-5 (1E-4 in the last round) and purging the alignment of possibly non-homologous sequence fragments after each round. PSIPRED2 is used for secondary structure prediction. The alignment is converted to a HMM and compared with a database of domains of known structure. These HMMs were derived in the same way as the query HMM from a set of representative SCOP3 sequences (maximum sequence identity 50%). Their secondary structure states are determined by DSSP. HHpred is based on the HHsearch4 software (http://protevo.eb.tuebingen.mpg.de/download/). HHsearch makes use of HMM-HMM comparison and employs a score that generalizes the log-odds score of HMM-sequence comparison. Secondary structure is compared between HMMs using specially derived substitution matrices for secondary structure states.

HHpred.3 is similar to HHpred.2 but uses intermediate profile search to construct the query alignments. It looks at the results of the last iteration of PSI-BLAST when building the query alignment for HHpred.2 and picks seed sequences with E-values between 1E-4 and 1. New PSI-BLAST searches are started with theses seeds and the alignments are merged if an HMM-HMM comparison indicates homology between the two.

In a preliminary analysis, two main limitations for HHpred.2/3 in the CASP/CAFASP benchmark were discovered: (1) We based our template selection on an outdated structure database (SCOP v1.65) that did not contain the best templates in the pdb for many targets. (2) The database of alignments was constructucted for homology detection purposes instead of structure prediction. For homology detection to be reliable, one requires high selectivity, i.e. clean, and therefore less diverse alignments, whereas for structure prediction benchmarks, one should better use highly diverse alignments which yield a higher sensitivity.


1. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
2. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* ***292***, 195-202.
3. Murzin,A.G., Brenner,S.E., Hubbard,T.J. & Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* ***247***, 536-540.
4. Söding,J. (2004) Protein homology detection by HMM-HMM comparison, submitted to *Bioinformatics*.


## **Hirst-Nottingham** - 18 models for 18 3D targets

### Great deluge algorithm in CASP6


Y. Bykov[1,2] , M. T. Oakley[2], E. K. Burke[1] and J. D. Hirst[2]

*1 - School of Computer Science & IT,*

*2- School of Chemistry, University of Nottingham, United Kingdom.*

yxb@cs.nott.ac.uk


All structures presented by our group were produced by in-house optimization software that employs a multiobjective local search method called the “Great Deluge” algorithm.[1] This technique has performed well on other optimization problems,[2] which motivated its application to protein structure prediction.

The prediction of the native state of a protein was formulated as a continuous optimization problem, where the three-dimensional conformation with minimum energy needed to be identified. The representation of proteins employed retains most of their geometrical properties, i.e. it simulates all atoms except apolar hydrogen atoms, in an extended atom representation. The conformational space of the model is explored by variation of the torsion angles of the backbone and side chains.

The “Great Deluge” local search is an iterative procedure, where at each step a new conformation is randomly selected from a set of candidates generated from the current conformation (its neighborhood). The chosen candidate is accepted as the new current conformation if it fits into an artificial feasible space, which is gradually reduced during the search. This mechanism, unlike the Monte Carlo method, makes the local search process highly controllable by the user. In particular, it allows improvements to the accuracy of prediction by regulating the processing time and exploring different areas of a multiobjective search space. To make the algorithm more effective, different neighborhood structures are explored with different priority, i.e. the rotation of main and side chains, small and large changes in angles, simultaneous modification of two torsion angles, etc. Special attention was paid to acceleration of the search. Using a “delta-evaluation” mechanism the algorithm does not recalculate the complete energy function of the candidate at every step, but only its difference from the current conformation. This feature provides an almost linear scaling of the

evaluation time with the length of a protein. The current version of the software can work with proteins up to 140 residues. With such a protein it evaluates around 1000 conformations per second (on a PC P4 3.2GHz) and is able to produce a result in 10-15 hours (generally a result could be produced after evaluating around 50 million conformations).

In the course of our research, most investigations were focused on improving the effectiveness of the optimization technique, rather than verification of the energy function. Currently, our algorithm operates with the following energy terms: a sum (for all pairs of atoms) of Lennard-Jones potentials, a sum of electrostatic potentials (between charged atoms), a sum of hydrogen-bond potentials (in donor-acceptor groups) and a function modelling the hydrophobic effect (involving side-chain carbon atoms). The total energy is calculated as a weighted sum of these four components. The initial formulae and parameters of the energy functions were taken from the CHARMM package.[3] However, we are currently studying the possibility of verifying the energy parameters using a higher level search (where the described algorithm is performed as a low-level procedure). A special technique (based on a linear programming method) was developed for dynamic tuning of the weights using known proteins. Thus, the energy parameters used for prediction of CASP6 targets were automatically tuned in order to provide the best fit to the known native states of several short proteins.

Before submitting to CASP6 web site, all predicted structures were post-processed using CHARMM. This involved energy minimisation with harmonic constraints applied to the positions of the backbone atoms.

Acknowledgments: This work was supported by the BBSRC (grant 42/BIO14458) and the EU/Framework 6 'BioPattern' Network of Excellence.

1. Dueck, G. (1993). New Optimization Heuristics. The Great Deluge Algorithm and the Record-to-Record Travel. *J. Comput. Phys.* **104**, 86-92.
2. Burke, E. K., Bykov, Y., Newall, J. P., Petrovic, S. (2004). A Time-predefined Local Search Approach to Exam Timetabling Problems. *IEE Trans.* **36**, 509-528.
3. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., Karplus, M. (1983). CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comp. Chem.* **4**, 187-212.

**Hmmspectr3** - 206 models for 64 3D targets
**Hmmspectr_fold** - 120 models for 63 3D targets

## Protein structure prediction using combination of Hidden Markov Models (HMM) based search and modeling

Y.V. Sharikov, L.F. Ten Eyck, I.F. Tsigelny
*University of California, San Diego, San Diego Supercomputer Center*
itsigeln@ucsd.edu

For CASP6 predictions we used the advanced version of HMMSPECTR system HMMSPECTR3 (http://hmm-spectr.sdsc.edu) The system is based on searching for the best alignments between the target primary sequence and members of Hidden Markov Models (HMMs) library of protein structural homologs[1]. As compared with the previous version of HMMSPECTR, we changed the concept of selection of final protein structure predictions using combination of the HMM library searches with high throughput target modeling. In the Hmmspectr3 group we constructed the final prediction using human experience with known homology modeling programs, in the Hmmspectr_fold group we presented the results of automated prediction with some human corrections.

All protein targets presented in CASP6 competition one can divide to three groups: Proteins having such sequences (sequence identities more than 27-30%) that their structural relatives are found easily using simple BLAST-like search (26 targets from the entire list). The second group contains 15 proteins with the sequence identity 15-26%. For their processing one can use a set of HMM libraries: TIGR[2], Pfam library[3], more powerful Superfamily1.65 library[4], our own HMM-SPECTR library[1]. The rest of proteins – 35 targets - are outside of these two groups. These proteins have sequence identities less than 15% to any known proteins having solved crystal structures. Such proteins either have sequence identities lower to their real relatives than to random proteins from the libraries sets, or have complex domain structures. HMMSPECTR3 (http://hmm-spectr.sdsc.edu) reflects such stratification of possible targets. It makes protein structure prediction only after making decision about a specific type of targets. We used SCOP[4] and CATH[5] classifications, HMMER 2.2.1[6] program package, and our program Original Structure Alignment Tool[7] for construction of the comprehensive HMM library. Constructed HMM library is available for users (ftp://ftp.sdsc.edu/pub/outgoing/sharikov). To include to the library recently published protein structures we used the following technique. Using the HMM libraries constructed from the existing members of classification groups we did a search for primary sequences of new members of

PDB database. The proteins that had scores corresponding to the specific classification groups of protein structures were added to these groups and those HMM sets were reconstructed adding these new proteins.

Process of protein structure prediction (fold recognition) can be defined as having the following steps: (1) Selection of HMM having the greatest score with a target sequence, (2) Association of a set of greatest score parent proteins with this HMM, (3) Alignment of the target and parent using HMM. Two proteins parents are usually associated with a specific HMM. The first—showing the greatest score from the entire PDB proteins set. The second—showing the 'closest pattern of recognition' by target—corresponding boundaries in HMM, score value, general pattern of alignment (dominant residues, gaps location, etc.). In the case of sufficient alignment length (usually more than 50% of a target length) second protein parent usually very well defines a predicted protein structure. Using various values of a 'gapmax' parameter[6] and a number of members in the structural alignment for a specific HMM (using our pre-compiled libraries) we can obtain different alignments parent-target and eventually different final predictions. The final selection is based on a set of final models (for two different parents minimum two different templates) and the final models are assessed using ProQ[8]. The computation complexity is growing sharply in the cases of small differences between HMM selections for a target. It usually happens with low total scores and low alignment lengths. In such cases we check not a one, but a set of closest HMMs with minimum two proteins associated to each of selected HMMs. These proteins usually have to show the greatest scores for a specific HMM inside various length intervals of alignments (with the step around 10% of the entire alignment length). Final selection is based on the ProQ scores of each of predicted structures. In the case when no structure shows a reliable score with ProQ we start the following algorithm based on domain separate estimates. The domain model is created using the greatest score parent proteins on each of sequence regions having maximum scores with specific HMMs. In this case ProQ assessment gives low scores and for the final selection is used HMM score and to some criteria.

Below we describe in more details an algorithm of the HMMSPECTR3 work in the case of low sequence identity. We use here the concepts of 'secondary HMMs' and simultaneous secondary structure prediction for high throughput automatic construction of prediction models for their further selection using scoring by protein structure quality tools like ProQ, Verify_3D, etc. The first step in forming of the 'secondary HMM' is formation of 30-40 pairwise alignments. Alignment of a parent to a target is done using existing HMMs from the library. Then we select 3-5 HMMs with the greatest scores for each 'sequence length category'. Here we align only the dominant (having the greatest scores) proteins. For the more precise prediction one can use not only dominant proteins but the sets of greatest score proteins. Then we construct multiple alignments where a target protein sequence is not gapped. The inserts residues in the protein sequences aligned to it are excluded. Then the secondary HMM is constructed and is used for the search within the PDB databank.
To the set of supporting libraries we include a file 'pdb4-3' containing: ID, scoreID, primary_sequence, secondary_structure_sequence, start_border, end_border. This file is prepared using primary sequences of PDB proteins divided to the pieces corresponding to the secondary structures sets HHHHHH, EEEEEE, CCCCCC.

After the nomination of possible 'parent' protein we use it as an initial template. Let us presume, for example, that a template is from the c.2.1.4. class of the SCOP classification. The pieces of the primary sequences (prepared as described above) of all proteins of this class are aligned to the target protein taking in consideration their primary and secondary structures and predicted secondary structure of the target. The best fitting protein is used as a 'parent' and its C-alpha trace is used for further entire protein structural model generation. The side chains of all residues that are directly aligned within the corresponding secondary structures are inserted to the model. The side chains of other residues are inserted when the corresponding residue is found in pairwise alignment of a member of c.2.1.4. class with the target protein within the corresponding secondary structure regions. This process is continued until the entire set of c.2.1.4 class is examined. If there are still residues that do not correspond to the target primary sequence this procedure is repeated with the higher set of SCOP classification, in this case c.2.1, c.2. Eventually we construct a model with the C-alpha trace corresponding to the initially chosen parent protein and side chains constructed on the base of the entire set of the class or subclass of SCOP. Created model sometimes still having some gaps and unresolved regions can be already assessed by the protein structure quality programs and used for checking if the initial parent selection was right. A number of such models is created in automated mode and are used for selection of the final prediction parent protein. This approach definitely has some drawbacks. There exists a possibility that a model having highest scores by the protein structure assessment programs would be far away from the real target protein. Proper checking points scoring method is in development. The other drawback is 'overpricing' of longer alignment *vs*. shorter alignments. Often it is not true that the longest alignments would be more corresponding to the target. There are some other problems. The main advantage of such an approach is its complete automation. This way using high performance computers one can check a number of possible hypotheses in a short period of time, change weights of parameters used and even adjust a prediction system for specific classes of proteins.

Table 1. illustrates some limitations of HMM-based methods. Column 'PDB closest relative' shows protein that had to be chosen as a best prediction. All targets included to this table are not predicted well enough. For the targets T0216 and T0228 the lengths of alignments to the 'closest relative' is around 25% of the target protein length. For the targets T0198, T0202, T0270 sequence identities are lower than 10%, for the targets T0223, T0205 too many gaps. The latest case prevent the effective use of HMM alignment of target-parent pair. Initially a correct target is selected by HMM in the first selection set, but then on the stage of final selection incorrect alignment because of large gap percentage lead to the selection of other models having higher scores, but incorrect parents.

| CASP | Answer | Length | PDB closest relative | Z-score | RMSD (A) | Seq. ident. (%) | Length aligned part | Gaps | Hmm score | Hmm length |
|---|---|---|---|---|---|---|---|---|---|---|
| T0198 | 1SUM | 235 | 1SUN | 5.9 | 2.3 | 7.8 | 103 | 12 | 2.8 | 72 |
| T0202 | 1SUW | 249 | 1QO0 | 5.0 | 2.8 | 9.2 | 98 | 16 | 15.7 | 68 |
| T0205 | 1VM0 | 130 | 1EXB | 4.2 | 2.6 | 10.9 | 64 | 57 | 64.2 | 86 |
| T0216 | 1VL4 | 447 | 1IMU | 4.1 | 3.7 | 9.5 | 84 | 24 | 0.7 | 42 |
| T0223 | 1VKW | 218 | 2BKJ | 5.0 | 2.3 | 14.6 | 123 | 53 | 59.1 | 164 |
| T0228 | 1VLP | 441 | 1LTD | 5.0 | 2.9 | 8.1 | 99 | 24 | 3.6 | 41 |
| T0270 | 1VDH | 249 | 1MLI | 4.7 | 2.8 | 5.7 | 87 | 15 | 3.8 | 76 |


1. Tsigelny,I., Sharikov,Y., Ten Eyck,L. (2002) Hidden Markov Models-based system (HMMSPECTR) for detecting structural homologies on the basis of sequential information. *Protein Eng.* **15**, 347-352.
2. http://tigrblast.tigr.org/web-hmm/
3. http://www.sanger.ac.uk/Software/Pfam/
4. Murzin,A.G. et al. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536-540.
5. Pearl,F.M.G, Lee,D., Bray,J.E, Sillitoe,I., et al. (2000) Assigning genomic sequences to CATH *Nucleic Acids Research* **28**, 277-282
6. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics* **14**, 755-763.
7. Kotlovyi,V, Tsigelny,I., Ten Eyck,L.F. (2002) A flexible method for structural alignment: Application to structure prediction assessments. In *Protein structure prediction: Bioinformatic Approach* (ed. I.F.Tsigelny), pp. 433-447. IUL La Jolla, CA.
8. Wallner,B, Fang,H, Elofsson,A. (2003) Automatic consensus-based fold recognition using Pcons, ProQ, & Pmodeller *Proteins*. **53** S6, 534-41
9. Luthy,R, Bowie,J.U., Eisenberg,D. (1992) Assessment of protein models with three-dimensional profiles. *Nature* **6364**, 83-5


## **HOGUE-DFP** - 40 models for 8 3D targets

### The Distributed Folding Project


H.J. Feldman[1], E. Garderman[1] and C.W.V. Hogue[1,2]

*[1] – The Blueprint Initiative, Mount Sinai Hospital, Toronto, Canada,*

*[2] – Department of Biochemistry, University of Toronto, Toronto, Canada*

chogue@blueprint.org


This team made use of Distributed Computing to make CASP predictions. Approximately 2000 volunteers around the world participated in the Distributed Folding Project (http://www.distributedfolding.org/), volunteering their spare CPU cycles to run our software client. To decide which targets to attempt, the CAFASP website was visited for each new target, and we looked at the 3D-Jury score[1]. Those which were below about 30 were considered 'difficult' and marked for prediction.

The distributed client used a modified version of our TRADES algorithm[2], incorporating secondary structure prediction from PsiPred[3] and performing probabilistic walks in Ramachandran space. Sidechains were placed probabilistically using Dunbrack's backbone dependent rotamer library[4]. All residues are chirally and sterically valid, having a minimum of non-hydrogen van der Waals collisions.

Approximately one billion structures were generated for each target using the Distributed Folding Project framework, in a time span of one week per target. An iterative approach was used to create 250 successive generations, such that each new generation is seeded with a conformational space map from the previous generation's best structure, as determined by a fitness score (see below). The first generation was large (30,000) and consisted of probabilistically generated structures. The later generations were much smaller (100) and all members were close in structure space to the seed structure for that generation. The result is a dynamics-like folding simulation as the structure travels through conformational space. Each participating CPU runs its own independent simulation of 250 generations.

Finally, from the pool of generated structures various statistics were collected including radius of gyration, exposed surface area, exposed hydrophobic surface area, and a fitness score – a modified version of a statistical residue-based potential[5] which also compares actual secondary structure content to

predicted content. This helps remove structures that are loopy and not protein-like. Additionally, to ensure only compact structures were retained, structures with radius of gyration greater than 120% * $2.59N^{0.346}$, where N is the number of residues in the protein, were all discarded. The best structures were chosen based on their fitness scores. The top 10 structures were visually inspected, and five chosen for submission.


1. Ginalski,K., Elofsson,A., Fischer,D. & Rychlewski,L. (2003). 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics* **19**, 1015-1018.
2. Feldman,H.J. & Hogue, C.W.V. (2000). A Fast Method to Sample Real Protein Conformational Space. *Proteins* **39,** 112-131.
3. Jones,D.T. (1999). Protein Secondary Structure Prediction Based on Position-Specific Scoring Matrices. *J. Mol. Biol.* **292**, 195-202.
4. Dunbrack,R.L.,Jr. & Cohen,F.E. (1997). Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci.* **6,** 1661-1681.
5. Bryant,S.H. & Lawrence,C.E. (1993). An Empirical Energy Function for Threading Protein Sequence through the Folding Motif. *Proteins* **16,** 92-112.


## **HOGUE-HOMTRAJ** (serv) - 105 models for 45 3D targets

### HomTraj: a fold recognition server using trajectory distributions


H.J. Feldman[1], K.A. Snyder[1], M.J. Dumontier[1,2], and C.W.V. Hogue[1,2]

[1] – *The Blueprint Initiative, Mount Sinai Hospital, Toronto, Canada,*
[2] – *Department of Biochemistry, University of Toronto, Toronto, Canada*
chogue@blueprint.org


We developed HomTraj, a powerful, fully-automated homology modeling and fold recognition server. Once a query is received, NCBI BLAST[1] (expect value cutoff 1e-20) is used to identify up to five highly homologous templates from the PDB. If this call fails, the Sequence Alignment and Modeling (SAMT2K) algorithm[2] is used to identify up to five structure templates, using a two-track Hidden Markov Model (HMM) – one track for sequence, and one for secondary structure. PsiPred[3] was used to predict secondary structure of the query for input to the HMM.

Next, using a modified version of our TRADES algorithm[4], the backbone alpha-carbon trajectory of each template was recorded, and a trajectory distribution built with the new sequence of the target. Each gapless stretch of alignment was replaced by a single fragment from the recorded trace. Where gaps occurred in the alignment, fragments were built to span the gaps. These fragments were created as follows: The "takeoff angles" were recorded starting from one residue prior to the gap and ending one residue following the gap, on the template structure. These consisted of six degrees of freedom - the distance between the start and end of the gap, two virtual Ca angles and three virtual Ca dihedrals. Then three atoms from each side of the gap were placed in space, according to the recorded takeoff angles. Alpha carbons required to fill the gap were then given arbitrary starting co-ordinates within the gap region, and a steepest descent energy minimization carried out. For the purposes of this minimization, the energy function consisted of virtual Ca bond length restraints, virtual Ca angles restraints, and a van der Waals term. The three anchoring atoms on either side of the gap were held fixed during the minimization. Finally, the resulting loop was incorporated as a fragment using its own Ca trace. Gaps may be shifted a few residues left or right in order to minimize the energy of the loop spanning the gap.

Roughly 50 structures were generated using the fragments obtained from the previous steps and our Foldtraj software, with bump checking slightly reduced. This process was repeated for each possible template found in the initial step. Using a modified version of a statistical residue-based potential[5] which we have termed "crease energy", the best structure generated from each template was chosen and submitted.

Domain Prediction with Armadillo

A separate server, on the same team, was used to predict domain boundaries for CASP. The servers presently do not talk to each other, but in the future HomTraj will normally do a domain prediction first, and then model each domain separately. The Armadillo Domain Prediction algorithm uses two amino acid indices that reflect the propensity of residues to be in domain linker regions. The first index, DLI (domain linker index), is constructed from the amino acid propensity of domain linkers from a non-redundant set of multi-domain protein structures from the Protein Data Bank. The second index, REI (residue entropy index), was normalized from previously reported sidechain entropy values[6]. Each was used to build a distribution of scores across multi-domain proteins. Sequences used for a prediction are turned into a numeric profile using the index values, which is subsequently smoothed using a low pass filter under a Discrete Fourier Transform. Domain linker predictions are made when the smoothed values pass a significance threshold. Domain linker predictions are not made between the 50 residues at the N- and C- terminus.

Domains are consecutively numbered and there is no current provision to attempt to predict non-contiguous domains.


1. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
2. Karplus,K., Karchin,R., Barrett,C., Tu,S., Cline,M., Diekhans,M., Grate,L., Casper,J. & Hughey,R. (2001). What is the value added by human intervention in protein structure prediction? *Proteins* **Suppl. 5,** 86-91.
3. Jones,D.T. (1999). Protein Secondary Structure Prediction Based on Position-Specific Scoring Matrices. *J. Mol. Biol.* **292**, 195-202.
4. Feldman,H.J. & Hogue, C.W.V. (2000). A Fast Method to Sample Real Protein Conformational Space. *Proteins* **39,** 112-131.
5. Bryant,S.H. & Lawrence,C.E. (1993). An Empirical Energy Function for Threading Protein Sequence through the Folding Motif. *Proteins* **16,** 92-112.
6. Galzitskaya,O.V. & Melnik,B,S. (2003). Prediction of protein domain boundaries from sequence alone. *Protein Sci.* **12,** 696-701.


## **HOGUE-STEIPE** - 177 models for 61 3D / 59 FN targets

### Folding with FOLDTRAJ


H.J. Feldman[1], K.A. Snyder [1], M.J. Dumontier[1], M.V. Brougham[1], B.A. Tuekam[1], F. Wu[1], B. Thiruvahindrapuram[2], B. Steipe[2] and C.W.V. Hogue[1,2]

[1] – *The Blueprint Initiative, Mount Sinai Hospital, Toronto, Canada,*
[2] – *Department of Biochemistry, University of Toronto, Toronto, Canada*
chogue@blueprint.org


For CASP6 we used a variety of prediction methods to try to predict as much information as possible about each protein target, using many of the tools Blueprint has developed over the past few years. Each of these is summarized below.

Homology Modeling

Our first step for manual 3D structure prediction was to look at the CAFASP website for each new target, and to look at the 3D-Jury score[1]. Those which were below about 40 were marked for *ab initio* prediction (see next section). For the remainder, the alignment to the best CAFASP hit was usually used as a starting point.

Next, using a modified version of our TRADES algorithm[2], the backbone alpha-carbon trajectory of the template was recorded, and a trajectory distribution built with the new sequence of the target. Each gapless stretch of alignment was replaced by a single fragment from the recorded trace. Where gaps occurred in the alignment, fragments were built to span the gaps. These fragments were created as follows: The "takeoff angles" were recorded starting from one residue prior to the gap and ending one residue following the gap, on the template structure. These consisted of six degrees of freedom - the distance between the start and end of the gap, two virtual Ca angles and three virtual Ca dihedrals. Then three atoms from each side of the gap were placed in space, according to the recorded takeoff angles. Alpha carbons required to fill the gap were then given arbitrary starting co-ordinates within the gap region, and a steepest descent energy minimization carried out. For the purposes of this minimization, the energy function consisted of virtual Ca bond length restraints, virtual Ca angles restraints, and a van der Waals term. The three anchoring atoms on either side of the gap were held fixed during the minimization. Finally, the resulting loop was incorporated as a fragment using its own Ca trace. Gaps may be shifted a few residues left or right in order to minimize the energy of the loop spanning the gap. In some cases, additional templates were used when their alignments spanned gaps in the primary alignment. In this case, fragments from the secondary template were used to bias loop-building, by adding torsional angle constraints to the energy minimization. Then up to 30 structures were generated using the fragments obtained from the previous steps and our Foldtraj software, with bump checking slightly reduced. Only the region of the target which was aligned to templates was modeled. Using a modified version of a statistical residue-based potential[3] which we have termed "crease energy", the best structure is chosen.

*Ab Initio* Prediction

For those targets which had a CAFASP score below 50, a different approach was taken using the *ab initio* mode of the TRADES software. First, the Armadillo consensus algorithm [Dumontier & Hogue, unpublished] or NCBI's Reverse Position-Specific BLAST[4] was used to split the target chain into several domains which were then treated as separate folding units. Next, 840 recurring structural motifs, ranging in length from 3 to 16 residues, were identified from a protein database. For each residue in the target domains, the probability that each motif is the correct fragment at that position was determined using Bayesian statistics [Steipe & Thiruvahindrapuram, unpublished]. Results from PSIPRED[5] performed on the target sequence further bias the motif probabilities. In building the structures with a modified

version of our TRADES algorithm[2], the motifs were used, according to their probability, to specify the phi-psi-omega angles for that length of the chain. Then up to 6,000,000 structures were generated using the fragments obtained from the previous steps and our FOLDTRAJ software. Using an atom-atom contact potential which includes a solvation term[6], the best structures were chosen. The data for the separate domains, if there was more than one, was then concatenated.

Function Prediction

Lastly, we were interested in testing some new methods we have developed for function/binding site prediction. This was done as follows. First, GO terms were found by using BLAST[7] on target sequences, and copying annotation from high-confidence hits (E-value below 0.001).

We then made use of BIND-BLAST (http://bind.ca/BINDBlast/) to look for interactions in the BIND database[8] consisting of a molecule similar to the CASP target. A human expert then examined the interaction record, took into consideration any information that was known about the CASP target, and then decided whether any information could be inferred about the target based on the demonstrated BIND interaction.

In a similar manner, SMID-BLAST (http://smid.blueprint.org/smid_blast.php) was used to identify potential small molecule binding sites on the target, based on known protein-small molecule interactions stored in the SMID database (manuscript in preparation). This allowed precise prediction of small molecule binding sites, based on the BLAST alignments. Again, a human expert took into consideration what was known about the target, as well as sequence conservation at the binding site and promiscuity of the small molecule, to help determine which hits were biologically interesting and not false positives.


1. Ginalski,K., Elofsson,A., Fischer,D. & Rychlewski,L. (2003). 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics* **19**, 1015-1018.
2. Feldman,H.J. & Hogue, C.W.V. (2000). A Fast Method to Sample Real Protein Conformational Space. *Proteins* **39,** 112-131.
3. Bryant,S.H. & Lawrence,C.E. (1993). An Empirical Energy Function for Threading Protein Sequence through the Folding Motif. *Proteins* **16,** 92-112.
4. Marchler-Bauer,A., Anderson,J.B., DeWeese-Scott,C., Fedorova,N.D., Geer,L.Y., He,S., Hurwitz,D.I., Jackson,J.D., Jacobs,A.R., Lanczycki,C.J., Liebert,C.A., Liu,C., Madej,T., Marchler,G.H., Mazumder,R., Nikolskaya,A.N., Panchenko,A.R., Rao,B.S., Shoemaker,B.A., Simonyan,V., Song,J.S., Thiessen,P.A., Vasudevan,S., Wang,Y., Yamashita,R.A., Yin,J.J. & Bryant,S.H. (2003). CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res.* **31,** 383-387.
5. Jones,D.T. (1999). Protein Secondary Structure Prediction Based on Position-Specific Scoring Matrices. *J. Mol. Biol.* **292**, 195-202.
6. McConkey,B.J., Sobolev,V. & Edelman,M. (2003) Discrimination of native protein structures using atom-atom contact scoring. *Proc. Natl. Acad. Sci. U S A* **100,** 3215-3220.
7. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
8. Bader,G.D., Betel,D. & Hogue,C.W.V. (2003). BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.* **31,** 248-250.


## **Honiglab** - 105 models for 46 3D / 28 FN targets

### Combining alignment sampling and *ab initio* methods for comparative modeling and fold recognition


Donald Petrey[1,2], Mark Fasnacht[1,2], Lucy Forrest[2], Mickey Kosloff[2], Shoshana Posy[1], Chris Tang[2], David Pincus[3], Xin Li[3], Jiang Zhu[1,2], Cinque Soto[2], Claudia Bertonati[2], Sharon Goldsmith-Fishman[2], Rich Friesner[3], and Barry Honig[1,2]

*[1] –Howard Hughes Medical Institute, [2] – Department of Biochemistry and Molecular Biophysics, Columbia University,*
*[3] – Department of Chemistry, Columbia University*
bh6@columbia.edu


In CASP6 we used recently developed software for the sampling and analysis of alignments along with previously developed model building software The new software includes two new programs for alignment sampling, dalign and gnoali, and new tools for alignment analysis and visualization incorporated into GRASP2[1]. Alignment sampling with dalign is accomplished by enumeration of "suboptimal" alignments. An alignment is "suboptimal" if it does not have the optimal similarity score that is generally the output of dynamic programming algorithms. The program gnoali uses a similar algorithm, but also incorporates a geometrically based gap penalty. The new tools incorporated into GRASP2 were used to display, analyze and combine alignments generated by these programs to multiple templates. Combining alignments by merging them into a structure alignment of the possible templates using GRASP2 was an important step for several of the targets. It both enhanced the sampling of alignments

over what could have been achieved using just a single template, and in some cases allowed us to recognize when one template was more appropriate for a particular region of the target sequence, even if that template had a lower overall similarity score. A guiding assumption of the methods used during CASP6 was that alignment sampling can be used to identify regions of a target structure that are likely to be different from the structure of the template. Such differences are a major source of error associated with all template-based prediction methods and can have an effect on all aspects of the problem, from template selection to model evaluation.

The hypothesis we tested was that variations in alignments generated by the methods described below would indicate the model building/refinement strategy that was appropriate for a specific region of the target sequence. These strategies included building multiple models using alternate alignments followed by model evaluation, *ab initio* structure prediction of short regions, composite models based on multiple templates or any combination of these three. Models were built using the programs NEST, SCAP[2] and LOOPY[3] developed in our group. We emphasize that the alignment sampling used here was accomplished with a single method. Preliminary analysis of our results suggests that alignment variability produced using these programs is similar to simply comparing alignments generated with different methods and that we were frequently able to generate the "correct" alignment. Recognizing it remains an unsolved and difficult problem, however. When it was determined that *ab initio* methods were necessary, we used methods developed in the Friesner group and new methods for model refinement that combine existing sampling algorithms with a generalized Born model of the solvent.

Detailed analysis of the use of the above methods for two specific CASP6 targets are provided in our FORCASP methods paper but the procedure used generally consisted of the following steps. Possible templates were identified using HMAP[4]. Once a suitable template was found, similar folds were identified and a multiple structure alignment of these folds was generated using GRASP2 and analyzed to determine conserved and variable regions. Alternate alignments were generated using four methods: 1) by varying the HMAP input parameters; 2) by aligning to different templates (all alignments to similar folds were considered, as long as a statistically significant e-value was produced); 3) by generating suboptimal alignments; and 4) by generating suboptimal alignments using a geometrically-based gap penalty.

The generation of suboptimal alignments in methods 3 and 4 above is a new feature incorporated into HMAP. An important new component of the algorithm used to generate suboptimal alignments is the ability to "mask" certain regions of the sequence if it is believed that variability in that region will be insignificant. For example, if the template and target sequence are only distantly related, it is usually unnecessary to consider alternate alignments in loop regions, since loops will most likely have a different conformation. Thus, the method for generating suboptimal alignments implemented in HMAP allows a user to consider only "significant" differences in alignment, such as shifts in beta strands, greatly increasing the efficiency of the alignment sampling. The use of geometrically-based gap penalties is also a new feature of HMAP. With this method, gaps in the alignment are assigned a value based on the geometric distance between the end points of the deleted region of the template.

Alignments generated by the various methods were compared and analyzed by merging them into a multiple structure alignment of the selected templates using GRASP2. Alignments generated by the CAFASP servers were also included. The primary purpose of this analysis is to determine what models should be built, i.e., identify regions of the alignments that vary "significantly". For example, if two alignments generated by the methods described above vary in a shift in a secondary structure element, models based on both alignments would be built and evaluated based on the methods described below. Regions where the alignments were highly variable and no consensus alignments was observed (loops and loop-helix-loop regions primarily) would be targeted for *ab initio* prediction. In addition, functional information relating to targets and templates was used to manually analyze and optimize alignments, specifically to make sure that functional residues in the template (identified from review of the literature) were aligned with residues conserved in the target family. The analysis of alignments and the determination of which models to construct was largely a manual process. An effort is underway to automate this process however.

Our strategy for ab-initio loop prediction is based on the methodology outlined in Jacobson et al.[5] In brief, the entire protocol can be divided into six stages and consists of iterative execution of the Protein Local Optimization Program (PLOP). Information about PLOP can be obtained from http://francisco.compchem.ucsf.edu/~jacobson/. The first stage corresponds to the generation of initial loop conformations. In the second and third stages, restricted sampling is performed on low-energy minima previously located. The fifth and sixth stages are identical to the second and third. The fourth stage is termed the "fixed stage", and is based on the assumption that fragments of the generated structures have reasonable RMSDs from the native. A priori, however, we do not know which fragments of our predictions are native-like. The "fixed stage" attempts to solve this problem by holding an increasing number of residues fixed on the termini and subsequently re-predicting the remainder of the loop. All conformations are scored via an effective potential

composed of an OPLS all-atom force field, the SGB model of polar solvation, a nonpolar estimator, and a number of correction terms.

For loop-helix-loop prediction we adopted the methodology outlined in Li et.al[6] which is an algorithm for performing sampling of helix position and orientations, along with the rebuilding of the flanking loops on both sides of the helix. The first step is the enumeration and screening of helix conformations. Two anchor points of the helix terminals are mapped onto a set of grid points within the bounding spheres with a specified cutoff radius. A set of positions of the helix are obtained by moving the helix as a rigid body using all six degrees of freedom. The positions are subject to filtering based on sterics and loop-length. The next step involves the clustering of helix positions using a K-means algorithm to remove redundant helix conformations. Step three and four involve flanking loop closure and refinement; the protocol is almost identical to the one outlined in the previous paragraph. Finally, the side chains on the whole loop-helix-loop region are subject to optimization and energy minimization. The loop-helix-loop sampling method has been incorporated into the PLOP package, and utilizes the same effective potential.

All models are then evaluated using a combination of methods. Comparison of the conformational free energy of the models after minimization using all-atom physical chemical energy functions with either the CHARMM22, OPLA-AA, or GROMOS force fields was carried out. Minimization was done using either a single dielectric constant of 10 or a generalized Born model as implemented in the GROMOS and TINKER packages. When minimization was performed using a single dielectric, solvent effects were treated with the FDBP/γ method. Simplified potentials were also used including a method developed in our group, as well as Verify-3D and D-Fire. When there was strong consensus among the methods favoring a particular model, that model was chosen for submission to CASP6. When there was no consensus, a decision based on a manual evaluation of the quality of the alignments was made.

We implemented a structure-based function prediction procedure, using software developed in the Honig group as well as publicly available servers, in addition to literature reviews. The procedure began with collating all existing function information (e.g. Pfam family, InterPro, GO) for the target. Sequence homologs were detected using BLAST and aligned with ClustalW to identify specific residues conserved within the target family. This information was used in our analysis of the target-template alignments, as described above. Once a model was built we performed electrostatic and phylogenetic analyses using the programs GRASP and ConSurf to identify putatively functional regions, for example a patch of charged residues or cluster of conserved residues. The results from ConSurf were visualized with Rasmol or GRASP2. Structural alignments between the models and templates were generated and visualized with GRASP2, to further analyze conservation between functional residues in the template and structurally corresponding residues in the target. In addition, the electrostatic features of the template and model were compared. Finally, functional information obtained from the literature was combined with sequence- and structure-based methods to identify putative functional sites. For example, if the literature suggested post-translational modification (but the site was not yet identified), the target sequence was submitted to the PredictProtein server to identify multiple possible sites for modification. For each of these sites, the degree of sequence and the structural location was compared to choose the most likely modification site.


1. Petrey,D. and Honig,B. (2003). GRASP2: Visualization, surface properties, and electrostatics of macromolecular structures and sequences. *Meth. Enz.* **374**, 492-509.
2. Xiang,Z. and Honig,B. (2001). Extending the accuracy limits of prediction for side-chain conformations.[erratum appears in *J. Mol. Biol.* 2001 Sep 14;312(2):419]. *J. Mol. Biol.* **311**, 421-430.
3. Xiang,Z., Soto,C.S. and Honig,B. (2002). Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 7432-7437.
4. Tang,C.L., et al. (2003). On the role of structural information in remote homology detection and sequence alignment: new methods using hybrid sequence profiles. *J. Mol. Biol.* **334**, 1043-1062.
5. Jacobson,M. et al. (2004). A hierarchical approach to all-atom loop prediction. *Proteins* **55**, 351-367.
6. Li,X., Jacobson,M.P. and Friesner,R.A. (2004). High-resolution prediction of protein helix positions and orientations. *Proteins* **55**, 368-382.

**HU** - 28 models for 19 3D targets

## Consensus over transitive PSI-Blast alignments

A. Heger[1], T. Hätinen[1], S. Mallick[1], S. Shkumatov[1], C.A. Wilton[1] and L. Holm[1,2]

*[1] – Institute of Biotechnology, [2] – Department of Genetics, University of Helsinki*

liisa.holm@helsinki.fi

A successful strategy for protein structure prediction relies on identifying homologous sequences with known structure. Many proteins have only remote relatives in the structure database which are difficult to detect by sequence-based methods. With the rapid growth of sequence databases, the chances of being able to link distant homologues by a series of more closely spaced intermediate sequences are growing, too. We have developed a method for transitive alignment that uses intermediate sequences as stepping stones to infer an alignment between distant homologues[1]. This method was implemented in the MF server, which we used for fold recognition. Alignments generated by the automatic method were then manually refined. The MF server is based on pre-processing the information in an all-against-all alignment library[2] to enable instantaneous access to an optimal transitive alignment between any two proteins, no matter how many intermediates separate them. It uses a single sequence as input, computes transitive alignments to known structures, and returns the highest scoring alignment as the fold prediction. We selected 'hard' comparative modeling cases for manual prediction. Conserved motifs were identified as anchor points and the alignment of intervening segments was optimized with respect to solvation preference[3] and backbone continuity.

1. Heger,A., Lappe,M. & Holm,L. (2004) Sensitive detection of very sparse sequence motifs. *J. Comp. Biol*., *in press*
2. Heger,A, Holm,L. (2003) Exhaustive enumeration of protein domain families. *J. Mol. Biol*. **328**, 749-767.
3. Holm,L., Sander,C. (1992) Evaluation of protein models by atomic solvation preference. *J. Mol. Biol*. **225**, 193-205

**Huber-Torda** - 242 models for 63 3D / 60 RR targets

## Probabilistic fragments, optimized substitution matrices and fold recognition

T. Huber[1], T. Lai[2], E. Mittag[2], J.B. Procter[2], H. Stehr[2], S. Mühlenmeister[2], B. Otto,[2] A.E. Torda[2]

*[1] – Dept of Mathematics, University of Queensland, Australia ,*

*[2] –Centre for Bioinformatics, University of Hamburg, Germany*

torda@zbh.uni-hamburg.de

Philosophy

In the hands of many groups, protein threading means some combination of structure- and sequence-based terms. This is true of the "wurst" server. The emphasis in much of this work has been to treat as much as possible as parameters for optimizing and to use numerical optimization to find parameters which produce the best alignments on some calibration set of proteins. The philosophy even went as far as building a completely new amino acid substitution matrix.

Structure-based terms

The structure based score term gives the log-odds probability of a set of 9 residues matching a structural fragment of length 9. The implementation uses a fragment library, but it is rather different to those in the literature. Normally, one would classify fragments based on structural properties and then collect sequence statistics for each class of fragment. In contrast, the fragment-based scores were built by collecting $10^5$ fragments and using a Bayesian classification to sort them based on continuous descriptors (structural properties) and discrete descriptors (sequence) simultaneously.[2,3] This has interesting consequences. For example two classes may be structurally similar (both β-strand), but one reflecting a hydrophobic environment and one alternating hydrophobic/hydrophilic.

Optimization of alignment parameters

A very general method was used to optimize parameters. A parameterization set was collected, containing pairs of similar structures of low sequence identity. Within each pair, A & B, the sequence of A was aligned to its partner, B, producing a model for A. This could be compared to the original structure of A and used as the basis for a cost function. The better the parameters, the better the alignment and the lower the cost function. This was summed over a set of

1.5 x $10^3$ protein pairs.[2] The cost function was used in a simplex optimization and could be applied to gap and other penalties as described below.

Substitution matrix and full score matrix

Sequence alignments were calculated by a dynamic programming algorithm applied to an alignment matrix. This, in turn, was built by combining matrices from structure- and sequence-based terms using some weighting. This weighting was also treated as a parameter to be optimized and even the 210 elements of an amino acid substitution matrix were treated as parameters to be optimized. This could be used to produce a matrix for sequence alignments[4], but in this work, one wants something adapted to the rest of the scoring terms. To this end, an optimization was carried out of all gap penalties, the substitution matrix and the weights of the different terms simultaneously. Rather than use amino acid sequences, profiles from psi-blast were used in both sequence- and structure-based terms.

Preliminary results suggest that the optimization philosophy is very good at producing good sequence to structure alignment machinery. For this CASP, it is likely that our parameterization set contained too many pairs with high sequence similarity and was not tuned to the more difficult cases where the structural terms are most important. It also produced a substitution matrix which was too highly tuned to sequences with a large number of close sequence homologues. Although too late for CASP, these problems have already been repaired in recent re-parameterizations. A remaining weakness was our poor ranking of models. For some targets, a good model was in often in the top 20 guesses, rather than first rank. We are delighted with this feature as it leaves some other property to be optimized before CASP7.


1. http://www.zbh.uni-hamburg/wurst
2. Torda,A.E., Procter,J.B. & Huber,T. (2004) Wurst: A protein threading server with a structural scoring function, sequence profiles and optimised substitution matrices. *Nucl. Acids Res.* **32**, W532-W535.
3. Cheeseman,P. & Stutz,J. Bayesian classification (autoclass): Theory and results, in Advances in knowledge discovery and data mining, U. Fayyad, et al., Editors. 1995, The AAAI Press: Menlo Park. p. 61-83.
4. Qian,B. & Goldstein,R.A. (2002) Optimization of a new score function for the generation of accurate alignments. *Proteins* **48**, 605-610.
5. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W., & Lipman,D.J. (1997) Gapped blast and psi-blast: A new generation of protein database search programs. *Nucl. Acids Res*. **25**, 3389-3402.


## **ISTZORAN** - 190 models for 64 DR targets

### Combining predictors for short and long disorder


K. Peng[1], S. Vucetic[1] and Z. Obradovic[1]

[1] – *Center for Information Science and Technology, Temple University*

zoran@ist.temple.edu


During the past few years we have been focused on improving predictions for intrinsically disordered regions longer than 30 residues. As a most recent effort, four neural-network-based predictors, VL3, VL3H, VL3P and VL3E, were developed with prediction accuracies ranging from 83% (VL3) to 86% (VL3E).[1] However, all these predictors performed considerably worse on disordered regions shorter than 30 consecutive residues[1]. Similar behavior was also observed during our participation in the previous CASP experiment: all three VL3 type predictors (VL3E not used) successfully predicted both long disordered regions in the target proteins with accuracy higher than 80%, but were less successful on short disordered regions.[2]

There are several reasons for such a performance. First, the window lengths for attribute construction ($W_{in}$) and post-filtering ($W_{out}$) were optimized for predicting long disordered regions. Second, the training data did not include disordered regions of 30 residues or shorter. Third, a detailed analysis revealed that short disordered regions exhibit significantly different amino acid compositions and are more similar to flexible ordered regions in terms of flexibility index, hydropathy and net charge. [3] A predictor trained using a set of short disordered regions (3-10 consecutive residues) achieved only 66% accuracy on long disorder regions[3].

To address this problem, we developed a two-level predictor (model 1) which at the first level consisted of two specialized predictors: (1) a long disorder predictor for disordered regions longer than 30 residues, and (2) a short disorder predictor for disordered regions of 30 residues or shorter. At the second level, a predictor was built to determine which of the two first-level predictors should be used at a given position. Ideally, the two specialized predictors should receive weights of 1/0 in long disordered regions, 0/1 in short disordered regions, and 0.5/0.5 in ordered regions.

The dataset used contained a total of 1,335 non-redundant (all with <25% sequence identity) protein sequences, including (1) 153 proteins from DisProt[4] v1.2 (with DP0069 removed) with 163 long disordered regions and 24 short ones, (2) 511 PDB chains with 673 (43 long and 630 short) disordered regions

defined as stretches of missing coordinates[3], (3) 290 completely ordered PDB chains with no missing coordinates[3], and (4) 381 PDB chains released after June 2003 with 24 long and 329 short disordered regions. In total there were 230 long disordered regions with 25,958 residues, 983 short disordered regions with 9,632 residues, and 354,169 ordered residues.

The long disorder predictor was built using the same 20 attributes used for VL3 predictor and the net charge / hydrophobicity ratio calculated over a moving window of length 41 ($W_{in}$ = 41) centered at a current position. For the short disorder predictor 52 attributes were calculated over a much smaller window of 15 ($W_{in}$ = 15), including amino acid frequencies, *K2*-entropy, averaged flexibility, net charge/hydrophobicty ratio, averaged PSI-BLAST profiles, averaged secondary structure predictions, and an additional one indicating if the current position was located within 7 residues from the N- or C- terminus.

The second-level predictor is a 2-class predictor whose output indicates if a given sequence position is more likely to belong to a long disordered region. For a given sequence position, its class label was assigned by following rules: (1) **0** if more than half of a short disordered region overlapped with the subsequence of length 61 centered at that position, (2) **1** if more than half of a long disordered region overlapped with the subsequence, and (3) **1** if more than half of a short and a long disordered regions both overlapped with the subsequence. If a sequence position could not be labeled, it would not be used in training of the second-level predictor. The attributes used were the same as those used for the short disorder predictor except that they were calculated over a larger window of 61.

All three predictors were built as logistic regression models on balanced datasets of 16,000 randomly selected examples. Principal component analysis (PCA) was performed to reduce dimensionality by keeping variance at 95%. The outputs of the long and short disorder predictors were filtered by moving averaging windows ($W_{out}$) of length 31 and 5 respectively, while the outputs of the second-level predictor and the composite predictor were not smoothed.

To estimate prediction accuracy, the 1335 sequences were randomly divided into two disjoint sets (75%:25%) and the first part was used for predictor training and the second part for predictor evaluation. This process was repeated for 30 times and means and standard deviations of the resulting accuracies were reported. In this way, the *per-chain* accuracies for the composite predictor were estimated as 79.1±2.6%, 75.5±2.7% and 83.3±0.5% on short disordered, long disordered and ordered regions, respectively. For the two specialized predictors for long disorder and short disorder, the corresponding accuracies were 50.1±3.6%, 76.5±4.2%, 85.1±0.9% and 81.5±2.1%, 66.7±3.5%, 82.4±0.5%, respectively.


1. Peng,K., Vucetic,S., Radivojac,P., Brown,C.J., Dunker,A.K. & Obradovic,Z. (2004). Optimizing Long Intrinsic Disorder Predictors with Protein Evolutionary Information. *J. Bioinformatics and Comput. Biol.* (in press).
2. Obradovic,Z., Peng,K., Vucetic,S., Radivojac,P., Brown,C. & Dunker,A.K, Prediction of Intrinsic Protein Disorder, *Proteins* **53(S6)**, 566-572.
3. Radivojac,P., Obradovic,Z., Smith,D.K., Zhu,G., Vucetic,S., Brown,C.J., Lawson,J.D. & Dunker,A.K. (2004). Protein Flexibility and Intrinsic Disorder, *Protein Sci.* **13(1)**, 71-80.
4. Vucetic,S., Obradovic,Z., Vacic,V., Radivojac,P., Peng,K., Iakoucheva,L.M., Lawson,J.D., Brown,C.J., Sikes,J.G., Newton,C. & Dunker,A.K. (2004). DisProt: A Database of Protein Disorder, *Bioinformatics* [August 13, Epub ahead of print].


## **IUPred** - 57 models for 56 DR targets

### Prediction of protein disorder based on the estimation of pairwise interaction energy


Zsuzsanna Dosztányi, Veronika Csizmók, Péter Tompa and István Simon

*Institute of Enzymology, Biological Research Center, Hungarian Academy of Science, Budapest, Hungary*

zsuzsa@enzim.hu


Datasets of protein disorder are rather limited in size and heterogeneous in terms of the type of disorder they cover. The disorder prediction in CASP is restricted to only one type of disorder. i.e. missing residues in X-ray strutures. Instead of specifically addressing this subtype of protein disorder, we took a more general approach which could also provide a simple model for the physical basis of protein disorder. The underlying assumption is that globular proteins are composed of amino acids which have the potential to form a large number of favorable interactions, whereas IUPs adopt no stable structure because their amino acid composition does not allow sufficient favorable interactions to form. Based on this assumption, the polypeptides encoding globular and disordered proteins can be distinguished.

With the structure in hand, the energy of a protein can be easily calculated. Using a coarse-grained approach, the calculated energy is the sum of pairwise interactions between amino acid pairs within a distance cutoff. The energy of contacts between different amino acids, expressed in the form of a 20 by 20 matrix, was calculated from the observed frequencies of amino acid pairs using the approach of Thomas and Dill[1] The summation of such energies, however, cannot be carried out for proteins whose structure is unknown or for intrinsically unstructured proteins. To overcome these limitations, we invented a novel method for approximating the total pair-wise interaction energy from the amino acid compostion only[2] Without considering the actual conformation, we rely on statistics collected from a database of globular proteins which is used to derive the parameters for the estimation of the energy.

This novel approach is validated by the good correlation of this estimated energy with the values calculated for known structures. When applied for disordered sequences, their predicted energy values was clearly shifted towards less favourable energies compared to globular proteins. This indicates that experimentally characterized disordered proteins have special amino acid compositions, which, independently of the actual sequence, do not allow the formation of favorable contacts expected for folded proteins. Thus, these proteins are rightly called *intrinsically* unstructured.

At the core of our prediction method, termed IUPred, is the approximation of the pairwise energy by means of the amino acid composition of the protein. By limiting the calculation to a predefined sequential neighborhood, it yields a position-specific score characteristic of the tendency of a given amino acid to fall into a structurally ordered or disordered region. This score was averaged of over a given window size and normalized to fall between 0 and 1. For the specific targets in CASP, the cutoff value for the sequential neighbourhood and the window size was optimized on a database of ordered and missing residues in PDB structures. Although the optimization of these parameters brougth some improvements in the prediction accuracy of missing residues, we do not expect our method to outperform some machine learning algorithms directly trained for finding missing residues in X-ray structures. The real strength of our approach becomes apparent for full length proteins or domain-sized fragments of disorder, when this method, relying on a simple physical model only, outperfomes existing methods[2], like DISPROT VL3H[3] or DISOPRED2[4].


*1.* Thomas,P.D. & Dill,K.A. (1996). An iterative method for extracting energy-like quantities from proteins structures. *Proc Natl Acad Sci USA.* **93**, 11628-11633.

*2.* Dosztányi,Zs., Csizmók,V. Tompa,P. & Simon,I. (2004) The pairwise energy content esimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *submitted*.

*3.* Obradovic,Z., Peng,K., Vucetic,S., Radivojac,P., Brown,C., & Dunker,A.K. (2003). Predicting intrinsic disorder from amino acid sequence. *Proteins* **53 (S6)**, 566-572**.**

*4.* Ward,J.J, Sodhi,J.S., McGuffin,L.J., Buxton,B.F. & Jones,D.T. (2004) Prediction and functional analysis of native disorder in proteins form the three kingdoms of life. *J. Mol. Biol*. **337**, *635-645.*


## **JIVE** - 14 models for 14 3D targets

### JIVE: Protein structure prediction by the assembly of local supersecondary structural motifs


David F. Burke and Tom L Blundell

*Department of Biochemistry, University of Cambridge,80 Tennis Court Road, Cambridge, CB2 1GA, United KingdomInstitution*

dave@cryst.bioc.cam.ac.uk


In the CASP6 experiment, models of proteins which had low confidence values across the CAFASP4 servers were selected to be modelled.

JIVE predicts the structure of small conjoint domains of proteins by the assembly of fragments of local supersecondary motifs. Homologous sequences were identified using PSI-BLAST[1]. Secondary structure prediction was performed locally using PHD[2] together with the predictions from the CAFASP4 server. The conformational class of supersecondary fragments were predicted using SLOOP[3-5] based on all combinations of predicted secondary structure. The *SLoop* database contains protein loops clustered into distinct classes based upon the similarity of the mainchain conformation of their bounding secondary structures and loop residues. Each loop class is defined by an amino acid consensus pattern, the local structural environment of the loop residues and the angle and distance between the vectors of the bounding secondary structures. Models were built using a Monte Carlo simulation, assembling fragments derived from the predicted supersecondary motifs for contiguous loops together with fragments derived from the secondary structure predictions. Unsuitable models were rejected based on excluded volume and a distance-dependent conditional probability function[6]. The generated structures were then visually inspected to aid selection of likely models.

1. Altschul,S.F, Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.*Nucleic Acids Res.* **25**(17), 389-402.
2. Rost,B., *et al*. (1994) PHD-an automatic mail server for protein secondary structure prediction.Comput Appl Biosci.**10**(1), 53-60
3. Donate, L.E., *et al.*(1996) Conformational analysis and clustering of short and medium size loops connecting regular secondary structures: a database for modeling and prediction. *Protein Sci.* **5**(12), 2600-16
4. Rufino,S.D. *et al* (1997) Predicting the conformational class of short and medium size loops connecting regular secondary structures: application to comparative modelling. *J Mol Biol.* **267**(2), 352-67.
5. Burke,D.F. e*t al.* (2001) Improved Loop prediction from sequence alone. *Protein Engineering* **14** (7), 473-478
6. Samudrala,R *et al.* (1998) An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J Mol Biol*. **275**(5), 895-916

## **Jones-UCL** - 251 models for 63 3D / 64 DR / 26 FN

### FRAGFOLD3, THREADER3 and DISOPRED2: improved methods for prediction of protein folds, disorder and function

M.I. Sadowski[1], J.D. Watson[2], J.S. Sodhi[1], J.J. Ward[1] & D.T. Jones[1]

*[1] – Bioinformatics Unit, Department of Computer Science, University College London, Gower St., London, WC1E 6BT, United Kingdom*

*[2] – EMBL Outstation - Hinxton, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom*

dtj@cs.ucl.ac.uk

THREADER 3.5 is the latest incarnation of our original program to implement threading[1] (D.T. Jones et al., Nature 358, 86-89, 1992) and although it now incorporates a number of new features (in particular the use of sequence profiles), and a set of alignment parameters optimized with a genetic algorithm, the overall components of the current implementation remain more or less unchanged since CASP2. THREADER 3.5 was used to predict targets which were not predicted with high confidence by mGenTHREADER[2] or nFOLD (as submitted to the server prediction section). However, in making full CASP6 submissions, we also considered other models obtained from our web servers, and our new model quality assessment method (MODCHECK) was used to evaluate an ensemble of structures in order to identify the model predicted to have the highest accuracy.

For CASP6 targets which we believed could not be reliably predicted using fold recognition methods, FRAGFOLD3 [3] was used to generate up to 5 structures. This approach to protein tertiary structure prediction is based on the assembly of recognized supersecondary structural fragments taken from highly resolved protein structures using a simulated annealing algorithm. FRAGFOLD3 differs from previous versions by making use of both fixed-length and supersecondary structural fragments, explicitly modeling side chains using a fast rotamer generation method, and an improved treatment of main chain hydrogen bonding using a simple Morse potential. Up to 1000 structures were generated for each target domain using a 100 CPU Beowulf cluster, and a simple rigid-body structural clustering algorithm used to select the models representing the largest clusters of conformations. Submitted predictions were made using little or no human intervention apart from initial domain assignment and preparation of input secondary structure and sequence alignment files.

For all targets (including CM and FR targets), regions of native disorder were predicted using DISOPRED2 [4-5]. DISOPRED2 is based on a reimplementation of DISOPRED using Support Vector Machines rather than neural networks.
Predictions of the functions of the structurally and functionally uncharacterised targets for the CASP6 experiment were made using a manual approach combining information from a variety of sequence and structure-based methods, along with literature searching and visual inspection of predicted structures. Sequence-based methods used were the standard sequence similarity searching tools BLAST and PSI-BLAST[6], InterPro[7] and CDD[8] searches, STRING[9] and ANAGRAM[10]. The newly-developed TopSite program[11] for identifying metal-binding sites in low-resolution structural models and the ProFunc[12] ensemble of structural analyses (incorporating searches against ligand- and DNA-binding templates, SSM fold matching, nest analysis and SiteSeer searches) were also applied to the best structural predictions generated for each sequence. Results were then carefully analysed with reference to the results of the structural predictions and published information on the protein families predicted.

*1.* Jones,D.T., Taylor,W.R. & Thornton,J.M. (1992) A new approach to protein fold recognition. *Nature* **358**, 86-89.
*2.* McGuffin,L.J. & Jones,D.T. (2003) Improvement of the GenTHREADER method for genomic fold recognition. *Bioinformatics* **19**, 874-881.

3. Jones,D.T. (1997) Successful ab initio prediction of the tertiary structure of NK-Lysin using multiple sequences and recognized supersecondary structural motifs. *Proteins* **S1**, 185-191.
4. Jones,D.T. & Ward,J.J. (2003) Prediction of disordered regions in proteins from position specific score matrices. *Proteins* **S6**, 573-578.
5. Ward,J.J., Sodhi,J S., McGuffin,L.J., Buxton,B.F., Jones,D.T. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* **337**, 635-645.
6. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
7. Mulder,N.J. *et al.* (2003). The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res*. **31**, 315-318.
8. Marchler-Bauer,A., Bryant,S.H. (2004) CD-Search: protein domain annotations on the fly. *Nucleic Acids Res* **32**,W327-331
9. von Mering,C., Huynen,M., Jaeggi,D., Schmidt,S., Bork,P., Snel,B. (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res*. **31**, 258-261.
10. Perez,A.J., Thode,G., Trelles,O. (2004) AnaGram: protein function assignment. *Bioinformatics* **20**, 291-292.
11. Sodhi,J.S., Bryson,K., McGuffin,L.J., Ward,J.J., Wernisch,L., Jones,D.T. (2004). Predicting metal binding sites in low resolution structural models. *J. Mol. Biol*. **342**, 307-320
12. Laskowski,R., Watson,J.D., Thornton,J.M. *et al., unpublished.*


## **Karypis** - 61 models for 5 3D / 56 RR targets

## Prediction of contact maps using support vector machines


Ying Zhao and George Karypis

*Department of Computer Science, University of Minnesota*

yzhao@cs.umn.edu, karypis@cs.umn.edu


The problem of contact map prediction can be stated as a classification problem. Given a set of proteins with known structures, contact residues and non-contact residues are separated as positive instances and negative instances. For each instance, various features are collected to capture useful information of the pair of residues, including amino acid content, physicochemical environment, secondary structures, evolutionary correlation, and other information that can discriminate contacts from non-contacts. Then, these feature vectors of both positive instances and negative instances are used as the input to a classification tool to learn a classifier (*i.e.*, predictor). Given a sequence with unknown structures, the resulting predictor classifies the pairs of residues of the sequence to be contacts and non-contacts based on their feature vectors. In our RR model for CASP6, we employed Support Vector Machines (SVMs) as the classification tool and collected various features based on primary sequences, multiple sequence alignments, predicted secondary structures, and correlated mutation analysis[1] to predict contacts between non-local residues (sequence separation between the two residues is larger than 6).

Data Preparation

The dataset we used in training and testing our predictors contains 170 proteins with known 3D structures from Protein Data Bank (PDB[4]). The proteins whose chains are not interrupted and contain no more than two domains were selected. The list of proteins was further reduced to only contain the proteins with pairwise sequence identity lower than 25%. To obtain multiple sequence alignments (MSAs), we first used PSI-BLAST to retrieve homologous sequences for each protein and only kept sequences with more than 20% and less than 80% sequence identity. Then, we used ClustalW[6] to generate the final MSAs of the target protein and its homologous sequences. The predicted secondary structures for each protein were obtained by using PSIPRED[5].

Features

For each pair of positions in a protein sequence, we identified five sets of features that capture different aspects of the amino acids and the two locations: sequence separation, sequence conservation, predicted secondary structures, sequence profiles, and correlated mutations analysis.

The sequence separation between a pair of positions is the distance between two positions in the sequence. The conservation of each position in the sequence was calculated based on how conserve the amino acids appearing at that position in the multiple sequence alignment.

For each pair of positions, we consider the predicted secondary structures of both the two positions and their neighboring positions. In particularly, for each residue and its predicted secondary structure, we used three values to represent whether it belongs to an alpha helix, beta strand or coil. If the residue belongs to one of the three secondary structures, we set the corresponding value to be 1, and 0 otherwise.

The use of sequence profiles, which are derived from a multiple sequence alignment of homologous sequences, has been shown to be able to improve the prediction of contact maps[2]. We adopted the three-neighborhood approach in Ref[2]. For a pair of positions and their neighboring positions, we calculated the

sequence profiles as the occurrence frequencies of all the possible amino acid pairs from the multiple sequence alignment. In addition to using amino acid pair frequencies to represent the profile, we also used twelve physicochemical vectors from AAindex[4] to describe the physicochemical environment around. Specifically, for each position, the average of one physicochemical property was calculated by averaging the physicochemical property values for all the amino acid that appeared at that position in the multiple sequence alignment.

The correlated mutations analysis (CMA) utilizes evolutionary information. In evolutionary times, the significance of non-local contacts is manifested in the observed conservation patterns and the covariation of amino acid residues in multiple sequence alignments of homologous proteins. Pairs of distant sequence positions that are proximal in three-dimensional space appear to be conserved or mutated in a correlated fashion, *i.e.,* the frequencies of particular amino acid appearances in one position are dependent on the amino acid residue in the other position. In principle, positions with high correlation coefficients, a quantitative measure of mutational covariance in families of homologous proteins, can be inferred to be proximal in 3D. Specifically, we used the ten first principal components that resulted from a principal component analysis on 142 physicochemical vectors in AAindex [4] as the quantitative measures to calculate the correlation coefficients between pairs of positions based on the multiple sequence alignment of the target sequence. In addition, we also calculated correlated mutations defined in [2], which also employed similar correlation coefficient measure, but used pairwise amino acid scoring matrix of McLachlan instead of physicochemical vectors.

SVM Training and Prediction of Contacts

Given a training set of feature vectors of all the position pairs from all the sequences, we used SVM$^{light}$ [3] with a linear kernel and the default *C* value to train the SVM model. Since there are much more non-contacts than contacts, we randomly sampled non-contact instances, so that the number of contact instances and the number of non-contact instances are the same approximately.

Given a target sequence, the input for our predictor is also a collection of feature vectors of all the position pairs of that sequence. The predictor will return a score for each instance. Since we assign contact to be the positive class and non-contact to be the negative class, the higher the score is, the more likely the pair of amino acids is in contact. Hence, the returned scores can be sorted into a list, from which the top pairs are predicted as contact points. In our RR model for CASP6, we set the total number of predicted contacts from the sorted score list to be the total number of amino acids of the target sequence divided by 2. Finally, local contacts (sequence separation between the two residues is less than or equal to 6) were predicted based on sequence separation and predicted secondary structures, and all the local contacts were added to the final contact set as well.


1. Zhao,Y. & Karypis,G. (2003). Prediction of contact maps using Support Vector Machines. *In Proc. of the 3rd IEEE International Symposium on Bioinformatics and Biomedical Engineering (BIBE 2003).* 26-33.
2. Fariselli,P., Olmea,O., Valencia,A. & Casadio,R. (2001). Prediction of contact maps with neural networks and correlated mutations. *Protein Eng.* **14**(11), 835–843.
3. Joachims,T. (1999). Making large-Scale SVM Learning Practical. *Advances in Kernel Methods - Support Vector Learning*. Schlkopf B. and Burges C. and Smola A. (ed.), MIT-Press.
4. Kawashima,S., Ogata,H. & Kanehisa,M. (1999). AAindex: Amino acid index database. *Nucleic Acids Research*. **27**, 368-269.
5. Berman,H.M., Bhat,T.N., Bourne,P.E., Feng,Z., Gilliland,G., Weissig,H. & Westbrook,J. (2000). The Protein Data Bank and the challenge of structural genomics. *Nature Structural Biology.* **7**, 957–959.
6. Jones,D.T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195-202.
7. Thompson,J.D., Higgins,D.G. & Gibson,T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*. **22**, 4673–4680.


## **Keasar** - 283 models for 58 3D targets

### Refinement of fold recognition models by optimization with cooperative potentials


N. Kalisman, A. Levi, E. Erez, K. Noy, and C. Keasar
*Department of Computer Science, Ben-Gurion Universiry, Israel*
keasar@cs.bgu.ac.il


Fold Recognition (FR) emerges as a successful and promising approach to protein structure prediction. However, FR models tend to be fragmented and to include non-physical inter-residue distances. Such models may not be very useful beyond the somewhat artificial context of prediction experiments like CASP. Thus, we believe that the refinement of FR models is a major challenge in current computational structure biology.

Specifically, we try to generate non-fragmented, all-atoms models that are as similar as possible to the FR models, and at the same time physically plausible. Both requirements can be formulated into a derivable potential and the problem then becomes an optimization task. We implement this approach in BEAUTIFY, a new program handling many aspects of protein structure prediction including loop building and energy based optimization. BEAUTIFY is based on MESHI our in-house software package for molecular structure modeling.

Similarity of the BEAUTIFY model to the original FR template is enforced by distance constrains extracted from the template. *a-priori* all the distances between Cα atoms in the FR model may serve as constrains. In general, however, not all these constrains can be satisfied simultaneously in a physically plausible model. The optimization is thus done in several runs. The less satisfied constrains are removed after each run.

Physical plausibility is enforced by knowledge-based energy terms extracted from a non-redundant set of high-resolution structures (based on ASTRAL[1)]. Bond, angle, plane, out-of-plane and Van-der-Waals terms result in correct local structure and resolve clashes. On a higher level, we try to achieve "protein-like" appearance of the models by using cooperative energy terms that involve a large set of atoms coupled in a non-linear way. While more complex than the other terms, all the cooperative energy terms are derivable and evaluated in a linear time.

The cooperative energy terms include:
1) Hydrogen bond pairs - This energy term assigns low energy values to HB pairs frequently observed in proteins, such as the characteristic patterns of beta sheets. HB pairs that never occur in proteins are concurrently penalized by high-energy values. Usage of this term was shown to enhance the formation of native-like alpha/beta structures[2].
2) Solvation - This energy term induced a native-like solvation environment around every atom by forcing a certain number of neighboring carbon atoms in its vicinity.
3) Torsion Pairs - Low energetic values were assigned to frequently occurring torsion pair conformations, such as the allowed regions of the Ramachandran plot or the chi1/chi2 of common side chain rotamers.

In the current round of CASP we tried to refine Cα-models extracted from the CAFASP4 site. Depending on the variability of the models submitted to CAFASP, we manually chose from one to five template models. If an educated guess could be made considering the position of some missing residues, their Cα-atoms were added manually. These models were fed to the program together with a secondary structure prediction (a consensus of PSIPRED[3] and SAM-T02[4]). The refinement was done in three steps. First, the Cα-model was completed and refined, than the other backbone atoms were added and finally those of side chains. In all stages missing atoms were initially assigned random positions and reasonable structures were obtained by direct energy minimization. The random positioning of the missing atoms made this process non-deterministic, and many alternative decoys could have been generated from each template. The number of decoys actually generated ranged from one to 4000, depending on protein size and availability of computing resources. The resulted decoys were clustered and low energy representatives of the major clusters were submitted to CASP.


1. Brenner,S.E., Koehl,P., Levitt,M. (2000). The Astral compendium for protein structure and sequence analysis. *Nucleic Acids Res*. **26**, 254-256.
2. Keasar,C, Levitt,M. (2003) A novel approach to decoy set generation: designing a physical energy function having local minima with native structure characteristics. *J Mol Biol*. **329**, 159-174.
3. Jones,D.T, (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**: 195-202.
4. Karplus K, Karchin R, Draper J, Casper J, Mandel-Gutfreund Y, Diekhans M, Hughey R. (2003) Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. *Proteins* **53,** Suppl. 6, 491-496.


## **KIAS** - 675 models for 64 3D / 64 DP / 64 RR

### Prediction of residue-residue contacts using correlated mutation and hydrophobic packing score


Mee Kyung Song, Keehyoung Joo and Jooyoung Lee*
*School of Computational Sciences, Korea Institute for Advanced Study*
*207-43 Cheongryangri-dong, Dongdaemun-gu, Seoul 130-722, Korea*
jlee@kias.re.kr


Pair-wise residue contacts are predicted using the information on residue covariation[1] and conservation.[2,3] The covariation is determined from correlated mutation and the conservation from hydrophobic packing score between two positions in multiply aligned sequences. The contacts are predicted by three different methods; with correlated mutation only, with hydrophobic packing

score only, and finally with a combination of two as was done by Olmea and Valencia.[4] A contact is assumed between two residues when the minimum heavy-atom distance between them is less than 4.5 Å. All short range contacts less than four residue sequence separation are excluded.

For target proteins, the multiple sequence alignment (MSA) is carried by PSI-BLAST, using the non-redundant protein sequence database, with default parameters and the maximum of three iterations. With selected sequences, the following filtering process is carried out; sequences containing gaps of more than 22% of the target sequence and those with sequence identity greater than 95% are removed. The remaining sequences are used for contact prediction.

The correlated mutation score $C_{ij}$ between residues $i$ and $j$ is calculated as described by Gobel *et al*.[1] Each position in the alignment is represented by the corresponding element in the McLachlan matrix.[5] Residue pairs are sorted by their average correlated mutation score to predict contacts.

The hydrophobic packing score $H_{ij}$ is calculated from the sequence conservation coupled with hydrophobicity data.[3] The sorted list of residue pairs by their average hydrophobic packing score is used for contact prediction.

For the combined method, the score function between residues $i$ and $j$ is defined as $f_{ij} = C_{ij} + w H_{ij}$, where $w$ is the relative weight of $H_{ij}$ with respect to $C_{ij}$. The value of $w$ is chosen so that the best performance is achieved for a set of 281 domains selected from the SCOP database 1.63. From the total of 49497 domains in the SCOP, a set of domains containing sequence identity no more than 10% to any of its members is constructed. Out of 457 such domains, small domains containing less than 50 residues as well as domains with mutated residues are removed. Domains with the total number of aligned sequences less than 15 are also excluded to reduce statistical errors to obtain the 281domains. The prediction accuracy is defined by the number of correct contacts divided by the total number of predicted contacts. The best accuracies achieved for the 281 domains are 17.4%, 20.9%, 23.5%, and 37.7% for the number of predicted contacts of $L/2$, $L/5$, $L/10$, and 1, respectively, $L$ being the length of domain. In CASP6, the parameter for $L/2$ is used to predict contacts.

For each target, we first perform domain prediction using PPRODO.[6] If the target is predicted as a single-domain protein, residue-residue contacts are predicted as described above. Otherwise, predictions are carried out separately for each domain. All three methods described above are employed to submit answers in CASP6.


1. Gobel,U., Sander,C., Schneider,R. & Valencia,A. (1994). Correlated mutations and residue contacts in proteins. *Proteins* **18**, 309-317.
2. Mumenthaler,C. & Braun,W. (1995). Predicting the helix packing of globular proteins by self-correcting distance geometry. *Protein Sci*. **4**, 863
3. Aszodi,A., Gradwell,M.J. & Taylor,W.R. (1995). Global fold determination from a small number of distance restraints. *J. Mol. Biol*. **251**, 308-326.
4. Olmea,O. & Valencia,A. (1997). Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Folding & Design* **2**, S25-S32.
5. McLachlan,A.D. (1971). Tests for comparing related amino acid sequences. *J. Mol. Biol.* ***61***, 409-424.
6. Sim,J., Kim,S-Y. & Lee,J. (2004). PPRODO: Prediction of PROtein Domain boundaries using Neural Networks. *submitted*.


## Tertiary structure prediction for comparative modeling, fold recognition and new fold targets in CASP6


Keehyoung Joo[1], Jejoong Yoo[1], Kyoungrim Lee[1], Hyung-Rae Kim[1], Seung-Yeon Kim[1], Mee Kyung Song[1], Ju-Beom Song[2], Sang Bub Lee[1,3], Sung Jong Lee[4], Jooyoung Lee[1*]

*[1]School of Computational Sciences, Korea Institute for Advanced Study*
*[2]Department of Chemistry, Kyungpook University, Korea;*
*[3]Department of Physics, Kyungpook University, Korea*
*[4]Department of Physics, Suwon University, Korea*
jlee@kias.re.kr


For blind prediction of 3D structures of CASP6 targets, we have developed a unified method that can be applied to all classes of targets, called CMCSA (Combined Modeling using Conformational Space Annealing). The CMCSA method is based on an energy function designed from the information on the radius of gyration, hydrophobicity, $C_\alpha$ - $C_\alpha$ contacts, restraints from templates, restraints from super-fragments, restraints for β-pairing, hydrogen bond rewarding and steric hindrance. The energy function is given as

$$E = w_{rg}E_{rg} + w_{hp}E_{hp} + w_{MJ}E_{MJ} + w_{rst}E_{rst} + w_{hb}E_{hb} + w_{sc}E_{sc} \qquad (1)$$

where $w$'s are the weights of energy components.

Conformations are constructed by assembling fragments generated from PREDICT[1]. The PREDICT provides the secondary structure information of target proteins, libraries of local structure fragments, and $C_\alpha$ - $C_\alpha$ restraints of super-fragments extracted from the PDB_SELECT_90 by fold recognition

developed by us. For fragment assembly, we have used PROFESY[2], which was successfully applied to new fold targets in CASP5. Conformational search was carried out by conformational space annealing (CSA) method [3].

A standard set of weights in eq. (1) is obtained by parameter optimization using "representative" proteins selected from the SCOP database. For targets without additional information, the standard weights are used. For targets with "sure" templates (homology and threading targets), a larger weight is assigned for the restraints from templates. Weights are varied depending on the secondary structures of all $\alpha$ proteins, all $\beta$ proteins, $\alpha/\beta$ proteins, and $\alpha+\beta$ proteins.

Methods for the design of the energy function

Each component of the energy function is designed as follows.

(i) $E_{rg}$: the component on the radius of gyration. The average value of radius of gyration of proteins of $N$ residues is $<R_g> = 2.2 \times N^{0.38}$. Thus, we set that the structure with its radius of gyration larger than the average value would have a high energy score. One simple choice of such an energy component is

$$E_{rg} = \max\{R_g - (2.2 \times N^{0.38} + 0.5), 0\} ,$$

where $R_g$ is the radius of gyration of the protein model.

(ii) $E_{hp}$: the component for hydrophobicity. Parameters for hydrophobicity are calculated from the mean $C_\alpha$ distance from the center of mass for each type of amino acid for proteins in ASTRAL 1.65. The energy component for hydrophobicity is designed so that it becomes smaller when hydrophobic residues are located at inner regions of a protein and hydrophilic residues are at outer regions. We define such an energy component as

$$E_{hp} = - \sum_i D_i H_i / R_g ,$$

where $D_i$ and $H_i$ are, respectively, the $C_\alpha$ distance from the centroid and the hydrophobicity of the $i$-th residue. This component therefore has a tendency to force hydrophobic residues to form a core inside a protein.

(iii) $E_{MJ}$: the Miyazawa and Jernigan type contact-energy component. Contact frequencies are calculated from the PDB_SELECT_90 database and the parameters for the contact matrix $A(R_i, R_j)$ are determined, where $R_i$ and $R_j$ are the residue types of the $i$-th and $j$-th residues respectively. Two residues are assumed to be in contact if their $C_\alpha$ distance is less than 7 Å. The energy component is defined as

$$E_{MJ} = - \sum_{i,j} A(R_i,R_j) , \quad \text{for } i\text{-}j > 4.$$

(iv) $E_{rst}$: $C_\alpha$- $C_\alpha$ restraints energy. The $C_\alpha$ - $C_\alpha$ distance restraints are generated from three sources, templates when available, super-fragments, and $\beta$-pairing. Super-fragments are contiguous fragments obtained from the fold recognition method, $\beta$-paring restraints are to ensure the paring of $\beta$-strands. The energy component based on $C_\alpha$ - $C_\alpha$ restraints is defined as

$$E_{rst} = \sum X_{min}(i,j) / d_{ij} ,$$

where $X_{min}(i,j)$ is the minimum value of the difference between $d_{ij}$ and all restraints for $i$ and $j$, $d_{ij}$ being the $C_\alpha$ distance between the $i$-th and $j$-th residues.

(v) $E_{hb}$: Hydrogen bond rewarding term. While modeling, two residues whose $C_\alpha$ distance lies between 2.6 Å and 3.6 Å with favorable bond directions get reward for gaining a hydrogen bond. The energy component is defined as

$$E_{hb} = -\sum_{ij} V_i \cdot V_j ,$$

where $V_i$ and $V_j$ represent two vectors forming a hydrogen bond.

(vi) $E_{sc} = E_{\alpha\alpha} + E_{\beta\beta} + E_{CC} + E_{NN} + E_{OO}$: Penalties for steric clashes. Modeling may cause steric clashes between two residues. We calculate pairwise distances between all backbone heavy atoms from PDB_SELECT_90 and find the minimum distances that are rarely allowed. For example, two $C_\alpha$'s rarely come closer to each other than 3.9 Å. For $C_\alpha$ - $C_\alpha$, $C_\beta$-$C_\beta$, $C$-$C$, $N$-$N$, and $O$-$O$, the minimum allowed distances, $d_{min}$, are 3.9 Å, 3.4 Å, 3.7 Å, 3.6 Å, and 2.8 Å respectively. With $d_{max} = (d_{min} + 5)$ Å, the component is defined as

$$E = \sum_{i,j} \{(d_{max} - d_{ij}) / (d_{max} - d_{min})\}^8 ,$$

where $d_{ij}$ is the distance between two backbone heavy atoms $i$ and $j$.

Procedure

The prediction procedure consists of the following four steps.

(i) *Prediction of secondary structure and construction of fragment libraries*. We employ PREDICT which is based on the nearest-neighbor method on the pattern space. The PREDICT generates sequence profiles using PSI-BLAST and defines the pattern for each residue. Each pattern is compared with those in the pattern database construted from PDB_SELECT_90, and 100 closest patterns to a query residue are selected to determine its secondary structure. In addition, for each residue, out of the 30 closest patterns, a fragment library of backbone dihedral angles containing 15 consecutive residues is constructed.

(ii) *Distance restraints from templates, super-fragments and β-pairing.* When templates with reasonable confidence are available, $C_\alpha$- $C_\alpha$ restraints for aligned parts are generated. In practice, we have used the results from the Meta Server[4]. In all cases, additional restraints are generated by analyzing the results of PREDICT. The PREDICT generates 100 nearest-neighbor patterns for each residue of a target sequence. A super-fragment is defined as a collection of contiguous residues along the target sequence where a particular protein in the PDB provides one of the 100 patterns. These super-fragments are sorted according to their residue lengths, and the top 100 of them provide $C_\alpha$ - $C_\alpha$ restraints among them. When PREDICT indicates that there are more than one $\beta$-strands, $C_\alpha$ - $C_\alpha$ restraints of all possible combinations of $\beta$-pairing are generated. Finally, all $C_\alpha$ - $C_\alpha$ restraints are put together in the energy term.

(iii) *Global optimization of the energy function by CSA.* In order to obtain a collection of diverse low-energy conformations, we apply CSA to the energy function of eq (1) where conformations are generated by fragment assembly. This is a variation of the PROFESY[2], a prediction method used for new fold targets in CASP5/6. Initial conformations are generated as follows. We randomly pick a fragment for each residue from its library. We then assemble these fragments in an order from N- to C-terminal by shifting one residue at a time. If a fragment does not join smoothly to the existing assembled structure, the current fragment will be discarded and a new one is selected from the corresponding library. Two fragments are assumed to join smoothly if they satisfy the constraints $|\varphi 1-\varphi 2|\leq 30°$ and $|\psi 1-\psi 2|\leq 30°$, or $|\varphi 1-\varphi 2|+|\psi 1-\psi 2|\leq 45°$. After conformations are generated, they are subsequently minimized by a local minimizer; one residue in the sequence is selected at random, and a fragment corresponding to the residue is selected from the library. If the replacement of the new fragment improves the energy score, the new conformation is kept and otherwise, the replacement is rejected. This procedure is repeated until the energy score does not improve any further. The conformational search is carried out by Conformational Space Annealing (CSA) method[3]. CSA provide us a bank of diverse conformations with low lying minima in the conformational space.

(iv) *Model selection*. Typically, a final CSA bank contains 100 conformations, which are grouped into five clusters by a K-means algorithm. The best conformation from each cluster is taken. The final five models are selected according to their scores.


1. Joo,K., Kim,I., Lee,J., Kim,S-Y., Lee,S. & Lee,J. (2004). Profile-Based Nearest Neighbor Method for Pattern Recognition. *J. Korean Phys. Soc*. **42**, 599-604.
2. Lee,J., Kim,S-Y., Joo,K., Kim,I. & Lee,J. (2004). Prediction of protein tertiary structure using PROFESY, a novel method based on fragment assembly and conformational space annealing. *Proteins* **56**, 704-714.
3. Lee,J., Scheraga,H.A. & Rackovsky,S. (1997). New optimization method for Conformational Space Annealing, *J. Comp. Chem*. **18**, 1222-1232.
4. Ginalski,K., Elofsson,A., Fischer,D. & Rychlewski,L. (2003). 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics* **31**, 3291-3292.
5. Sali,A. & Blundell,T.L. (1993). Comparative protein modeling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779-815.
6. Grotthuss,M., Pas,J., Wyrwicz,L., Ginalski,K. & Rychlewski,L. (2003). Application of 3D-Jury, GRDB, and Veryfy3D in Fold Recognition. *Proteins* **53**, 418-423.


## **KIST-CHI** - 127 models for 40 3D targets

### Prediction of protein structure using homology modeling technique


Myung Whan Chi and Jin Su Song

*Korea Institute of Science and Technology, Cheongryang, Seoul, Korea*

zambo@kist.re.kr


The homology modeling technique predicts the three-dimensional structure of a given protein sequence (target) based on an alignment of the protein to one or more homologous proteins (templates) of known structure. This technique becomes more and more important because the structural information from x-ray crystallographic or NMR results is increased. In this study we carried out conventional homology modeling approaches. The target protein was aligned with the templates which selected using PSI-BLAST[1] search against PDB (Protein Data Bank) database. Then, the template coordinates of aligned regions were transferred to target. The coordinates of the regions which not aligned were given using small fragment amino acid library. If the matched amino acid fragment was not found, the conformation search was carried out. The energy minimization and molecular dynamics simulation were performed to refine the model structure.


1. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.


## **KIST-CHOI** - 220 models for 60 3D targets

### Protein structure prediction by fold recognition


Han Su Choi[1], Young Sun Kim[1], Jin Kak Lee[1, 2] and Chan No Yoon[1]

*[1] - Korea Institute of Science and Technology, Cheongryang, Seoul, Korea*

*[2] - Nanormics, Inc. 10-57 Hawolgokdong, Sungbukku, Seoul, Korea*

chs@kist.re.kr

For identification of template structure we used PSI-BLAST[1] against the non-redundant sequence database and fold recognition program. Fold recognition program searches sequence structure alignment using predicted secondary structure (PSI-PRED[2]), solvent accessibility, and sequence property. It is designed so that the best performance is achieved at twilight zone with low sequence identity. From sequence structure alignment, we carried out the target protein modeling by MODELLER[3] and side-chain modeling was followed by SCWRL[4] program. Then, energy minimization and molecular dynamics simulation were performed to refine the target structure.


1. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
2. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195-202.
3. Sali,A. & Blundell,T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779-815.
4. Dunbrack, R.L., Jr., Karplus, M. A. (1993) backbone dependent rotamer library for proteins: application to sidechain prediction. *J. Mol. Biol.* **230**, 543-571.


## **KOLINSKI_BUJNICKI** - 303 models for 64 3D targets

### Generalized protein structure prediction based on combination of fold-recognition with *de novo* folding and evaluation of models


A. Koliński[1] and J.M. Bujnicki[2]

*[1] - Faculty of Chemistry, Warsaw University, Pasteura 1, 02-093 Warsaw, Poland, [2] – International Institute of Molecular and Cell Biology, Trojdena 4, 02-109 Warsaw, Poland*

kolinski@chem.uw.edu.pl, iamb@genesilico.pl


To predict the tertiary structure of full-length sequences of all targets in CASP6, regardless of their potential category (from easy homology modeling to apparent new folds) we used a novel combination of two very different approaches that performed quite well in different categories in CASP5: the "FRankenstein's Monster" approach for comparative modeling (CM) based on recombination of Fold-Recognition (FR) models[1], and a new implementation of a Replica Exchange Monte Carlo method for protein structure prediction de novo or with restraints[2;3].

Sequences of all CASP6 targets were processed by the GeneSilico structure prediction meta server, which is a gateway to a variety of third-party methods for prediction of protein primary and secondary structure, solvent accessibility, and protein fold-recognition (see http://genesilico.pl/meta/[4], for links to all methods). Fold-recognition (FR) alignments were compared, evaluated, and ranked by PCONS and structures corresponding for up to 5 most frequently reported folds were selected for further analysis. For each candidate fold, the alignments between the target sequence and the structures of selected templates were used as a starting point for modeling using the "FRankenstein's monster" approach[1]. Best models obtained (1-15 models for each fold) were used to derive spatial restraints from those amino acids that exhibited VERIFY3D[5] score > 0.2. Additional restraints were derived from CAFASP models submitted by third-party fully automated servers for de novo structure prediction. Secondary structure restraints were derived from the consensus of methods implemented in the GeneSilico meta server[4]. Tertiary restraints derived from the FR and de novo models as well as secondary restraints derived from the consensus prediction guided the Replica Exchange Monte Carlo (REMC) folding simulation using a new high-resolution reduced lattice CABS model[2;3]. The CABS model employs a lattice-confined Cα representation of the main chain backbone, with 800 possible orientations of the Cα-Cα virtual bonds. The side-chains are off-lattice. The force-field of the CABS model contains several components that mimic averaged interactions derived from statistical analysis of the structural regularities seen in globular proteins. The effect of the solvent is treated in an implicit manner as an averaged contribution to the interaction of the side chains (see www.biocomp.chem.uw.edu.pl and ref.[2;3] for details). Results of the CABS simulations were subject to the average linkage hierarchical clustering algorithm with the distance root-mean-square separation as a measure of structures similarity. For each cluster its centroid was calculated and a full atom model rebuilt. Selection of final models was based on the combination of objective criteria, such as the energy of the models and the size of the respective clusters, and subjective visual analysis to reject models that exhibited features unlike to appear in real proteins, such as atypical angles of strands in beta-sheets or rare handedness of connections between elements of secondary structure.


1. Kosinski, J., Cymerman, I.A., Feder, M., Kurowski, M.A., Sasin, J.M., and Bujnicki, J.M. (2003). A "FRankenstein's monster" approach to comparative modeling: merging the finest fragments of Fold-Recognition

models and iterative model refinement aided by 3D structure evaluation. *Proteins* **53 Suppl 6**, 369-379.
2. Boniecki, M., Rotkiewicz, P., Skolnick, J., and Kolinski, A. (2003). Protein fragment reconstruction using various modeling techniques. *J Comput Aided Mol Des* **17**, 725-738.
3. Kolinski, A. (2004). Protein modeling and structure prediction with a reduced representation. *Acta Biochim Pol* **51**, 349-371.
4. Kurowski, M.A., and Bujnicki, J.M. (2003). GeneSilico protein structure prediction meta-server. *Nucleic Acids Res* **31**, 3305-3307.
5. Luthy, R., Bowie, J.U., and Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature* **356**, 83-85.

## **LANL_PFIG** - 16 models for 16 FN targets

### Nearest neighbor categorization for function prediction

K. Verspoor, J. Cohn, S. Mniszewski and C. Joslyn
*Los Alamos National Laboratory*
verspoor@lanl.gov

We present the methods utilized in a system aimed at predicting the function of CASP targets, as represented by a node in the Gene Ontology[2]. The strategy we follow is to (1) identify close neighbors of a target sequence in sequence space, (2) collect the Gene Ontology nodes associated with these neighbors in a curated data set (Swiss-Prot), and (3) categorize the collection of Gene Ontology nodes based on their distribution in the Gene Ontology structure, utilizing a technology called the Gene Ontology Categorizer[4]. The resulting set of Gene Ontology nodes is interpreted as the most representative nodes for the function of the original target sequence.

To identify close neighbors of a target sequence, we performed a PSI-BLAST (Position-Specific Iterated BLAST)[1] search on the target against the NCBI NR database, with 5 iterations. We used the default e-value threshold of 10.

Once the nearest neighbors in sequence space of the target sequence have been identified, we must collect the Gene Ontology (GO) nodes associated with these sequences. To achieve this, we first obtain the Swiss-Prot identifiers annotated to each PSI-BLAST match using a parsed listing of the NR database headers. Then, using the SIB/EBI Swiss-Prot to GO mappings, we find all of the Gene Ontology nodes related to the corresponding proteins. Finally, we build a weighted collection of Gene Ontology nodes, where each node in the collection is given a weight according to the PSI-BLAST e-value. Since several near neighbors of the original target sequence may map to the same Gene Ontology nodes, the collection we build can have redundancy. In this case, each occurrence of a Gene Ontology node will be weighted individually according to its source.

This collection of weighted Gene Ontology nodes becomes the input query to a categorization technology called the Gene Ontology Categorizer (GOC)[4]. This technology aims to identify a set of nodes in the Gene Ontology which best summarize or categorize a given list of input nodes. The technology is based on a view of bio-ontologies as combinatorially structured databases rather than facilities for logical inference, and draws on the discrete mathematics of finite partially ordered sets (posets) to develop data representations and algorithms appropriate for the Gene Ontology. Briefly (for more detail, see references 4,6), after identifying the set of input nodes in Gene Ontology space, GOC traverses the structure of the Gene Ontology, percolating hits upwards, and calculating scores for each Gene Ontology node. GOC then returns a rank-ordered list of Gene Ontology nodes representing cluster heads. In the end, this provides an assessment of which nodes best cover the input set.

We consider the set of cluster heads returned by GOC to be indicative of the function of the collection of nearest neighbors of the target sequence, and hence indicative of the function of the target sequence itself. These are returned as the predictions for the functions of the target sequence (subject to thresholding of the GOC results) and submitted to the CASP assessors.

The GOC system has many parameters that need to be specified in order to run effectively. To establish appropriate parameter settings for the CASP predictions, we created a “gold standard” test set of protein sequences for which mappings to Gene Ontology nodes were known. The test set consisted of the distinct set of Swiss-Prot sequences associated with entries in the 1.65 version of the SCOP dataset[5] through Protein Data Bank[2] annotations. This set was filtered to include only those sequences that had mappings in Swiss-Prot to the Gene Ontology, resulting in 774 test sequences. We measured precision and recall results for the GO function predictions over this test set for different parameter values, making sure to eliminate a PSI-BLAST match to the original sequence itself to avoid biasing the GOC analysis. For the system used to generate the submitted results for the CASP targets, we selected the parameter values which corresponded to the best empirical balance of precision and recall over the test set.

*Acknowledgements:* This work was sponsored by the Department of Energy under contract W-7405-ENG-36 to the University of California. We would like

to thank the Los Alamos National Laboratory Protein Function Inference Group for their contributions to this work.

1. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
2. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G. Bhat,T.N., Weissig,H., Shindyalov,I.N., Bourne,P.E. (2000). The Protein Data Bank. *Nucleic Acids Res.* **28,** 235-242.
3. The Gene Ontology Cosortium (2000). Gene Ontology: Tool For the Unification of Biology, *Nature Genetics* **25** (1), 25-29.
4. Joslyn,C., Mniszewski,S., Fulmer,A., Heaton,G. (2004). The Gene Ontology Categorizer. *Bioinformatics* **20**, Suppl. 1, i169-i177.
5. Murzin,A.G., Brenner,S.E., Hubbard,T., Chothia,C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536-540.
6. Verspoor,K., Cohn,J., Joslyn,C., Mniszewski,S., Rechtsteiner,A., Rocha,L.M., Simas,T. (2004). Protein Annotation as Term Categorization in the Gene Ontology using Word Proximity Networks. To appear in *BMC Bioinformatics*.

## **LOOPP** (serv) - 320 models for 64 3D targets

### Fold recognition by machine learning approach

Jian Qiu[1], Jaroslaw Pillardy[2], Tamara Galor[2], Craig B. Lowe[1], Leonid Meyerguz[1] and Ron Elber[1]

[1]*Department of Computer Science, Cornell University, Ithaca NY 14853,*
[2]*Cornell Theory Center, Cornell University, Ithaca NY 14853*
ron@cs.cornell.edu

LOOPP (Learning, Observing, and Outputting Protein Patterns) is a program to build structural models based on information from related proteins. LOOPP emerged from our earlier studies of folding potentials using Mathematical Programming approaches[1,2,3]. We have trained numerous scoring functions/ energies that evaluate the fitness of a sequence to a structure. To fully test and appreciate the capacity of the newly developed potentials we developed a prediction algorithm around these potentials. The first version of the algorithm[4] was based primarily on matching sequences to structures. Since then we have extended and enhanced the algorithm by including numerous similarity measures that are going beyond the single feature of sequence-to-structure matching.

Roughly, the similarity measures/features are divided as follows. We consider general properties: sequence similarity, sequence-to-structure matching, secondary structure fitness, exposed surface area, (we use the secondary structure and exposed surface area prediction program Sable[5]), and matching to the sequence profile of the probe and target sequence families. Each of these properties is examined in multiple ways. We compute the raw score, the difference of the native score from the score of the reverse native sequence, and the Z score. We also compute a special threading energy[1] and a Z score of that special energy according to the alignment of the current feature. Since some of these measures are expensive to compute in large-scale predictions of protein structures, we divide the calculation into three steps. In the coarse level only similarity measures that can be computed rapidly are taken into account, and that excludes the calculations of the Z scores. The remaining scores are combined to a single similarity measure that is used to pick 50 top candidates from our database of structures.

The top 50 candidates are evaluated with the expensive scores. Those include (but not limited to) the Z scores. Other expensive features include the build-up of atomically detailed models (generated with the MODELLER program of Andrej Sali[6]) and the assessment of this model using novel energy functions.
The cheap and the expensive measures are finally combined to a single similarity measure that ranks the models and provides the structures for the top 20 models.

1. Tobi,D., & Elber,R. (2000). Distance dependent, pair potential for protein folding: Results from linear optimization. *Proteins, Structure Function and Genetics* **41**, 40-16.
2. Meller,J., Elber,R. (2001). Linear Optimization and a double Statistical Filter for protein threading protocols *Proteins, Structure, Function and Genetics* **45**,241-261.
3. Teodorescu,O., Galor,T., Pillardy,J., Elber,R., (2004) Enriching the sequence substitution matrix by structural information. *Proteins, Structure, Function and Genetics*, **54**, 41-48.
4. Frary,A., Nesbitt,C., Frary,F., Grandillo,S., van der Knaap,E., Cong,B., Liu,J., Meller,J., Elber,R., Alpert,K.B., Tanksley,S.D. (2000) Cloning, Transgenic Expression and Function of fw2.2: a Quantitative Trait Locus Key to the Evolution of Tomato Fruit. *Science* **289**, 85-88.

5. Adamczak,R., Porollo,A., Meller,J. (2004) Accurate Prediction of Solvent Accessibility Using Neural Networks Based Regression. *Proteins: Structure, Function and Bioinformatics*, **56**, 753-67.
6. Marti-Renom,M.A., Stuart,A., Fiser,A., Sánchez,R., Melo,F., Sali.,A. (2000) Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 291-325.

## **LOOPP_Manual** - 258 models for 56 3D targets

### An alignment algorithm using residue type, secondary structure and solvent accessibility information to enhance accuracy of structural models

Jian Qiu
*Department of Computer Science, Cornell University, Ithaca NY 14853*
jianq@cs.cornell.edu

LOOPP_manual is a modeling procedure that picks candidates for structural templates by the LOOPP server http://ser-loopp.tc.cornell.edu/cbsu/loopp.htm and manually refines them to atomically detailed structures. In the first step, the top 20 templates returned by the LOOPP[1-4] server and top hits from PSI_BLAST[5] are combined to make up the set of template candidates.

In the second step, an alignment between the target and a candidate template is generated with a novel substitution matrix that is based on three complementing statistical potentials derived from structural alignments. These three potentials include a residue-residue substitution matrix, a residue type vs. secondary structure-surface area type matrix, and a predicted secondary structure_surface area type vs. actual secondary structure_surface area type matrix. Secondary structure and surface area predictions are computed with program SABLE[6] from Prof. Jaroslaw Meller's group, and actual secondary structure and surface area values of the templates are computed with program DSSP[7]. To complete the parameters required for generating the optimal alignment between the probe sequence and the template using dynamic programming, a position-specific gap penalty scheme was developed from structural alignments. This scheme includes residue-type-dependent gap penalty, secondary structure-surface area-dependent gap penalty and SABLE prediction-dependent gap penalty.

In the third step, an atomic model is generated based on each of the alignments with the program MODELLER[8]. The resulting atomically detailed models are evaluated, and a series of different scores are computed from the models, including atomic potential-based scores, the correlation between the actual secondary structures and exposed surface areas of the models and the SABLE-predicted values, sequence similarity between the query and the templates, and LOOPP scores. Visual inspections complemented with these scores are used to select the best models for submission.

1. Tobi,D. and Elber,R. (2000) Distance dependent, pair potential for protein folding: Results from linear optimization. *Proteins* **41**, 40-16.
2. Meller,J. and Elber,R. (2001) Linear Optimization and a double Statistical Filter for protein threading protocols. *Proteins* **45**, 241-261.
3. Teodorescu,O., Galor,T., Pillardy,J. and Elber,R. (2004) Enriching the sequence substitution matrix by structural information. *Proteins* **54**, 41-48.
4. Frary,A., Nesbitt,C., Grandillo,S., van der Knaap,E., Cong,B., Liu,J., Meller,J., Elber,R., Alpert,K.B., Tanksley,S.D. (2000) Cloning, Transgenic Expression and Function of fw2.2: a Quantitative Trait Locus Key to the Evolution of Tomato Fruit. *Science* **289**, 85-88.
5. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W., Lipman,D.J., (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389-3402.
6. Adamczak,R., Porollo,A. and Meller,J. (2004) Accurate Prediction of Solvent Accessibility Using Neural Networks Based Regression. *Proteins* **56**(4), 753-67.
7. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. (1983) *Biopolymers* **22**(12), 2577-637.
8. Marti-Renom,M.A., Stuart,A., Fiser,A., Sánchez,R., Melo,F., Sali,A. (2000) Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 291-325.

## **LTB_Warsaw** - 259 models for 62 3D targets

### Multitemplate modeling by a hierarchy of high-resolution lattice folding and all-atom refinement

D. Gront, A. Oleksy, P. Klein and A. Koliński
*Warsaw University, Faculty of Chemistry*
*Pasteura 1, 02-093 Warsaw, Poland*
dgront@chem.uw.edu.pl, kolinski@chem.uw.edu.pl

Our method starts from a number of molecular templates generated by threading metaservers. These templates provide a large set of distance restraints which guide folding using a reduced representation of protein conformational space. After clustering of folding results the final models are refined and ranked using all-atom force field and explicit solvent.

At the first step the threading models (20 top scoring templates) from bioinfo.pl metaserver[1] were compared to each other using structural pairwise alignment. In the cases of good consensus between various servers all templates were used as a source of distance restraints for a single folding simulations and the reduced models of templates used as a set of replicas for the Replica Exchange Monte Carlo Simulations using CABS[2,3] reduced-space modeling tool. In the cases of divergent results the structures from metaserver were clustered according to the crmd distance between them and the length of consensus alignment. Then, each cluster of templates provided a set of distance restraints for separate series of simulations. Additional restraints were derived from strongly predicted consensus secondary structure for regular fragments of structure (helices and beta sheets).

Large sets of distinct protein structures resulting from the CABS lattice simulations were then subject to a clustering procedure. Average linkage hierarchical clustering algorithm was employed with drmsd as the measure of the distance between structures. Cluster's centroids (averaging step in the clustering procedure) were computed via average distance maps. Finally 5-7 clusters were manually selected, according to the cluster size, average energy of its members and average distance dispersion (as a measure of the density of a cluster).

Starting from the alpha carbon trace for a cluster's centroid a full atom model was build using Pulchra algorithm[4]. Full atom models were then subject to long Molecular Dynamics simulations using the Amber[5] force field and an explicit solvent model. In the cases of "easy" CM/FR targets the MD simulations were limited to few steps and calculations of the all-atom energy) The lowest energy conformations were selected from the MD trajectories and subsequently optimized using conjugent gradient method. The resulting models were ranked according to their final all-atom energies and sent to the CASP server.

The method could be easily automated, provided a set of strict criteria for cluster selection is defined.


1. Ginalski K, Elofsson A, Fischer D & Rychlewski L. (2003) 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics.* **19** 1015-1023.
2. Kolinski, A. (2004) Protein modeling and structure prediction with a reduced representation. *Acta Biochim Pol*. **51**, 349-371.
3. Force field and other supplementary files for CABS model could be find on http://www.biocomp.chem.uw.edu.pl
4. Feig, M., Rotkiewicz, P., Kolinski, A., Skolnick, J., & Brooks, C. L.(2000) Accurate reconstruction of all-atom protein representations from side-chain-based low-resolution models. *Proteins* **41**, 86-97.
5. Pearlman, D.A., Case, D.A., Caldwell, J.W., Ross, W.R., Cheatham, III, T.E., DeBolt, S., Ferguson, D., Seibel, G., & Kollman, P. (1995) AMBER, a computer program for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to elucidate the structures and energies of molecules. *Comp. Phys. Commun.* **91,** 1-41.


## **Luethy** - 71 models for 64 3D / 7 FN targets

### Iterative sequence profile searches using a hardware accelerated Smith-Waterman algorithm


Roland Luethy
*TimeLogic Corp*
luethy@timelogic.com


Overview

The method described here is based on a iterative profile approach, similar to PSI-BLAST[1], but using a rigorous Smith-Waterman sequence database search step on a DeCypher hardware accelerator[2]. We have determined that this method is generally more sensitive than PSI-BLAST. In the first step, a profile was built from the target sequence and the sequences in the nonredundant protein database using the iterative profile method. The resulting profile was then used to scan sequences from the ASTRAL database[3] for high scoring sequences. The highest scoring PDB structure[4] was then used as the template to model the target. If the alignment of profile and PDB structure covered less than 60% of the target sequence, the target sequence was divided into subsequences, which were used to train profiles and build structure models.

Construction of profiles

Profiles were constructed in the same fashion as PSI-BLAST[1]: First a multiple sequence alignment was made from the hits of the previous run or from a single sequence search for the first iteration. The sequences with P-scores below 0.02 were aligned pairwise against the query sequence. These alignments were subsequently combined into a multiple sequence alignment using the initial

query sequence as an anchor. Sequences with pairwise identities greater than 94% were then removed from the alignment. Sequence weights were assigned using the position-based weighting method introduced by Henikoff[5]. Finally the position dependent scores were calculated as the natural log of $R_i$ using the following equation[1]:

$$R_i = \frac{\alpha(f_i / P_i) + \beta(\sum_j f_j r_{ij})}{\alpha + \beta}$$

where $f_i$ is the weighted observed frequency of amino acid *i* at the alignment position under consideration, $P_i$ is the frequency of amino acid *i* in the SWISS-PROT database, $\alpha$ is the average number of different amino acids per alignment position, $\beta$ is a pseudo count constant set to 9, $f_j$ are the weighted frequencies of all amino acids at the given alignment position and $r_{ij}$ are estimated ratios of frequency with which amino acids *i* and *j* are aligned. The values for $r_{ij}$ were estimated from BLOSUM62 substitution matrix S [6] with the formula $e^{0.316Sij}$.
Iterations were terminated when no new sequences were added to the alignment or after five iterations.

Model construction

First, all coordinates from the best scoring PDB structures were copied using the profile alignment as the guide. Gaps in the alignment where filled by finding overlapping short fragments from a database of PDB structures. Following this, the missing side-chain atoms were copied from the closest five-residue fragment from PDB with the identical middle residue. The structure was then minimized using TINKER[7] using the steepest descent method and a stepwise protocol that kept all C-alpha atoms fixed in the first step, those from the template were kept fixed in the second step and finally all atoms were allowed to move in the last step.

Conclusion

The method used here represents an improvement over PSI-BLAST with respect to sensitivity and speed.


1. Schäffer,A.A., Aravind,L., Madden,T.L., Shavirin,S., Spouge,J.L., Wolf,Y.I., Koonin,E.V., & Altschul,S.F. (2001). Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Research* **29**, 2994-3005.
2. www.timelogic.com (2003).
3. Chandonia,J.M., Walker,N.S., Lo Conte,L., Koehl,P., Levitt,M., & Brenner,S. E. (2002). ASTRAL compendium enhancements. *Nucleic Acids Res* **30**, 260-263.
4. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N., & Bourne,P.E. (2000). The Protein Data Bank. *Nucleic Acids Research* **28**, 235-242.
5. Henikoff,S. & Henikoff, J.G. 1994). Position-based sequence weights. *Journal Molecular Biology* **243**, 574-8.
6. Henikoff,S. & Henikoff,J.G. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* **89**, 10915-10919 (1992).
7. Ren, P. & Ponder, J. W. (2002). Consistent treatment of inter- and intramolecular polarization in molecular mechanics calculations. *J Comput Chem* **23**, 1497-506.


## **Luo** - 268 models for 54 3D targets

### Consistent scoring with AMBER/PB energy function


M.J. Hsieh and R. Luo
*Department of Molecular Biology and Biochemistry*
*University of California, Irvine, CA 92697*
rluo@uci.edu


Protein structure prediction at atomic detail, an important aspect of the protein folding problem, remains one of the fundamental unsolved problems in the field of computational molecular biology. There are primarily two classes of prediction methods for protein three-dimensional structure: comparative and ab initio predictions. No matter what method is taken, the final stage of protein structure prediction usually involves ranking or evaluating many protein models with a scoring function, an algorithm that gives a score for input structures to their fitness, that are used to judge the models likelihood of being the native structure, or at least of being close to the native.

There are two classes of scoring functions: knowledge-based and physics-based approaches[1,2]. The two scoring functions are constructed from very different starting points. Knowledge-based approaches are derived from distributions of experiment structural data. Physics-based approaches assume that the protein potential energy function can be broken down into terms of bond stretching, angle bending, torsional and nonbonded interactions. These parameters are then fitted to high-level ab initio quantum mechanical calculations and small molecule thermodynamic/spectroscopy data.

We have developed a physics-based scoring function (termed AMBER/PB)[3] based on an efficient Poisson-Boltzmann (PB) implicit solvent[4-6] and a refined

AMBER force field[7]. The accuracy in the PB treatment of the electrostatic interactions and the scalability of the particle-mesh treatment of long-range electrostatics make the scoring function well suited for targets up to protein domain boundaries[6]. In addition, the efficiency in the PB solvent[4] allows us to use the scoring function directly during the minimization phase before ranking, making it possible to develop a refinement method that directly applies the scoring function during sampling.

The scoring function for protein structure prediction has been analyzed with several widely used all-atom decoy sets. Testing on chosen decoy sets shows that the scoring function, designed to consider detailed chemical environments, is able to consistently discriminate all 62 native crystal structures after considering the heteroatom groups, disulfide bonds, and crystal packing effects that are not included in the decoy structures. When NMR structures are considered in the testing, the scoring function is able to discriminate 8 out of 10 targets.[3] In the more challenging test of selecting near-native structures, the scoring function also performs very well: for the majority of the targets studied, the scoring function is able to select decoys that are close to the corresponding native structures as evaluated by ranking numbers and backbone CαRMSD.[3] Various important components of the scoring function have also been studied to understand their discriminative contributions towards the rankings of native and near-native structures. It was found that neither the non-polar solvation energy as modelled by the SA model nor a higher protein dielectric constant improve its discriminative power. The terms remained to be improved are related to 1-4 interactions. We found that the most troublesome term is the large and highly fluctuating 1-4 electrostatics term, but not the torsion-angle term.[3]

To blind-test our scoring function in CASP6, we have taken initial all-atom models from two different sources: (1) all-atom models built in-house based on alignments deposited at the CAFASP4 prediction site, and (2) all-atom models deposited at the CASP6 prediction site. These models are then minimized in the AMBER/PB scoring function before initial ranking is performed. The top 10 models are then further refined in simulated annealing with the scoring function and re-ranked to select the final top 5 models for submission.


1. Moult,J. (1997) Database Potentials and Molecular Mechanics Force Fields. *Current Opinion in Structural Biology* **7**,194-199.
2. Lazaridis,T. & Karplus,M. (2000) Effective Energy Functions for Protein Structure Prediction. *Current Opinion in Structural Biology* **10**,139-145.
3. Hsieh,M.J & Luo,R. (2004) Physical scoring function based on AMBER force field and Poisson-Boltzmann implicit solvent for protein structure prediction. *Proteins* **56**, 475-486.
4. Honig,B. & Nicholls, A. (1995) Classical Electrostatics in Biology and Chemistry, *Science* **268**, 1144-1149.
5. Luo,R., David,L. & Gilson,M.K. (2002) Accelerated Poisson-Boltzmann calculations for static and dynamic systems. *J. Comp. Chem.* **23**, 1244-1253.
6. Lu,Q. & Luo,R. (2003) A Poisson-Boltzmann dynamics method with nonperiodic boundary condition, *J. Chem. Phys*. **119**, 11035-11047.
7. Lu,Q. & Luo,R. (2004) In Prep.


## **MacCallum** - 128 models for 64 3D / 64 RR targets

### Meta-server model ranking using predicted contact maps


R.M. MacCallum, B. Wallner and A. Elofsson

*Stockholm Bioinformatics Center, Stockholm University, Sweden*

maccallr@sbc.su.se


As described in more detail in the SBC group abstract (Wallner, et al.), we have made full-atom models using alignments taken from the bioinfo.pl metaserver[1] for all CASP6 targets and their homologues (if submitted). Various scoring schemes and energy calculations were applied to the models and the results were browsed via HTML tables (now at http://www.sbc.su.se/~arne/casp6). A new experimental score based on predicted contact maps was developed during the early stages of this prediction season. Encouragingly, the contact-based score seems to correlate with other measures, such as ProQ[2], and 3D-JURY[3], though we have not yet looked in detail at the (possibly trivial) reasons behind this. In the following, we describe the calculation and use of the contact prediction-based score.

We used contact predictions from our own approach[4], which were also submitted under the group name MacCallum in the RR category. All predicted contacts are separated by 24 or more residues, and for this purpose we take the *L*/2 most confident contacting pairs (*L* is the length of the target). For each all-atom model we then calculate two quantities:

1. the fraction of the predicted contact pairs that are actually present in the model – this is denoted *c*, and is a measure of coverage.
2. the mean C-beta to C-beta (C-alpha for glycine) distance between all predicted contact pairs *in the model* – this is denoted *d.*

We expect therefore to see a smaller mean distance, *d*, in the models which agree with our contact predictions (which we hope are correct). At the same time we don't want too many predicted contact pairs to be absent from the model.

Normalisation of *d* is required because it is quite strongly dependent on the length of the target. Starting with a plot of *d* against *L* for a set of SCOP domains, we derived two functions which approximated to the upper and lower limits of the distribution of *d* for any given *L*. These functions are as follows:

$$\text{lower}(L) = 3 * \log(L + 26) - \log(28)$$
$$\text{upper}(L) = 12.19 + (L - 12.72)^{0.5}$$

Then the normalised distance, $d_{norm}$, is calculated as:

$$d_{norm} = (d - \text{lower}(L)) / (\text{upper}(L) - \text{lower}(L))$$

Now we have two scores *c* and $d_{norm}$, both ranging from zero to 1 which we can plot on the "*x*" and "*y*" axes respectively. An ideal prediction would be found at the bottom right corner of this plot. In order to produce a single score from the twoscores, we calculate the Euclidean distance from the ideal (1,0) and subtract this from one:

$$\text{contact\_score} = 1 - \text{sqrt}((c - 1)^2 + d_{norm}^2)$$

Models derived from the servers that feed the meta-server are then ranked using the contact score alone. In most cases, only rank-1 models were used, but for some hard targets it seemed worth risking a non-rank-1 model if it had a much higher contact score. Some additional judgments were made based on the consensus of SCOP superfamilies, energy scores and the overall loopiness/knottedness of models. In general however, no more than about 15 minutes was spent on each target.

If this approach does provide an advantage, it is expected to be best for the more remote targets where alignment quality is poor, alignments may be partial, and fold assignment is not at all obvious. One possible limitation of this approach stems from problems with contact prediction itself; namely that most predicted contacts are rather short-range. Therefore the contact score will generally be higher for models with low contact order (fewer long-range contacts). This may be an issue with target T0279, where circularisation seemed to be an issue.


1. Bujnicki,J.M. Elofsson,A. Fischer,D. Rychlewski,L. (2001) Structure Prediction Meta Server *Protein Science* Nov, **10(11)**, 2354-62
2. Wallner,B. & Elofsson,A. (2003) Can correct protein models be identified? *Protein Science* May, **12(5)**, 1073-86
3. Ginalski,K. & Rychlewski,L. (2003) Detection of reliable and unexpected protein fold predictions using 3D-Jury. *Nucleic Acids Res.* **31(13)**, 3291-2.
4. MacCallum,R.M. (2004) Striped sheets and protein contact prediction. *Bioinformatics* **20** Suppl 1, I224-I231.


## **MacCallum** - 128 models for 64 3D / 64 RR targets
## **GPCPRED** (serv) - 63 models for 63 RR targets

### Contact map prediction from PSI-BLAST profile windows


R.M. MacCallum

*Stockholm Bioinformatics Center, Stockholm University, Sweden.*

maccallr@sbc.su.se


As previously described[1], we developed a simple approach to visualise sequence profile information on 3D protein structures. This involves clustering sequence profile windows (from proteins of known structure) using Kohonen's self organising map; then colouring the residues in a 3D protein viewer according to cluster identity. Due to the nature of the self organising map, neighbouring clusters have similar properties and are therefore assigned similar colours. Visual inspection of protein domains identified regularities in the colouring of beta-sheets. Parallel sheets often exhibit parallel striping of colour sequences, and neighbouring strand pairs in anti-parallel sheets occasionally showed reversed colour patterns. To test the generality of these observations, the transformed sequence profile information (residue colours) was used as the sole input (plus sequence separation) to a contact prediction algorithm. The results were surprisingly good and the prediction accuracy is expected to be equivalent to existing methods, even though it does not use any information about correlated mutations.

The target sequence of *L* residues is run through the default PSIPRED[2] version 2.3 scripts to produce a PSI-BLAST[3] ".mtx" text file containing the position specific scoring matrix of *L* columns by 21 rows. The rows correspond to the 20 amino acids and a mystery value, presumably related to indels. A total of *L*

overlapping windows of length *w* are extracted from the matrix, using zeroes to pad at each end. Each window (a *w* by 21 matrix) is mapped to a discrete position on a pre-trained self-organising map (SOM) which, in this work, is a 3D grid of 6 x 6 x 6 nodes. Note that the dimensionality reduction is substantial, particularly for larger windows (e.g. 15x21=315 reduced to 3). The 3D map coordinates can be converted into an RGB colour for visualisation or used as input to the prediction algorithm. Thus, a *L* residue sequence can be converted into a *L* by 3 matrix, for various sizes of window, *w*.

The "manual" RR predictions submitted by this group (MacCallum) are in fact produced with no manual intervention and are based on input transformations (see previous paragraph) using windows of size 1, 5, 9 and 15. The prediction algorithm is centred around the calculation of distances for pairs of residues *i* and *j*. Not all pairs are considered, first a subset of residues are selected using a filter function which takes the four input matrices, the residue position *i* and the sequence length *L* as input. The best-scoring *L*/5 residues are then passed to the pairwise distance calculation function, which takes the same inputs as the filter function, plus another residue index, *j*. Finally, the best scoring pairs (lowest distance) are considered as contacting residues. Typically one would select the best *L*/2, *L*/5 or *L*/10 for comparison with other methods.

How are the filter function and pairwise distance function implemented? Their internals are optimised using a type of evolutionary computing called genetic programming (GP). This is a population based search algorithm. Initially, individuals in the population each contain a random version of the two functions described above. The allowable expressions and operators are rigidly defined in a "grammar". A helper function is provided to facilitate the calculation of "colour pattern distances" between short parallel and anti-parallel segments of the input matrices. An individual is evaluated by applying the functions to the contact prediction problem on a periodically resampled set of 100 SCOP domains. The *L*/10 accuracy (fraction of predicted contacts that are real contacts; C-beta C-beta < 8.0Å) is used as the "fitness measure" to decide which individuals should reproduce and which should die. After some considerable amount of computation time, the accuracy of the predictors on the training set and an unseen test set is reasonably good (27% for *L*/10 predictions).

The results in the paper[1] are presented for a single individual picked from one of the 20 parallel evolving populations. In order to hedge our bets, the predictors used in CASP and the web-based service combine the results of a number of predictors sampled from these populations. The consensus method is relatively simple. Each predictor produces an *L* by *L* matrix of *i*,*j* distances, which are then ranked 1 (closest distance) to N (furthest distance), giving an *L* by *L* matrix of distance ranks. The final contact predictions are simply the residue pairs with the lowest mean ranks (averaged over different predictors).

The GPCPRED automated server uses a slightly different approach, with window sizes of just 1 and 15, and a different GP implementation, which works on the entire *L* x *L* matrix (there is no "filter" function).


1. MacCallum,R.M. (2004) Striped sheets and protein contact prediction. *Bioinformatics* **20** Suppl 1, I224-I231.
2. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195-202.
3. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.


## **MCON** - 63 models for 63 3D targets

### Selecting models with a meta-MQAP


Y. Azaria
*Center of Excellence in Bioinformatics, Buffalo, NY.*
azariaya@bioinformatics.buffalo.edu


An MQAP (Model Quality Assessment Program) is a program that receives as input a predicted 3D-model and returns a single number that represents its quality. An MQAP can do any computation, as long as the only input is a single model. The quality assessment is performed on the predicted model only, without any prior knowledge of the native structure itself. Traditionally, MQAPs correspond to programs that evaluate the "energy" of a model using some potential.

Six MQAPs (Solvx, Modcheck, Bala, ProQ, Verify3D and Prosa) that participated in the CAFASP-MQAP experiment (http://www.cs.bgu.ac.il/~dfischer/CAFASP4) plus a virtual MQAP developed in-house were combined to create an MQAP-consensus program. The MQAP-consensus program simply adds the z-scores of the individual MQAPs to produce a combined MQAP-consensus score. For CASP, the MQAP-consensus was applied to all the rank-1 predictions of the full-atom-generating CAFASP servers and the model with the highest sum-of-z-scores was submitted.

Notice that the above procedure is different from that applied by the MQAP-CONSENSUS of the CAFASP-MQAP experiment in that the latter considered all the full-atom models of the servers, and not just the rank-1 models.

MQAP-consensus is not a predictor: it is simply a “meta-selector”. The goal was to evaluate how successful a simple meta-MQAP is in selecting the best models from the rank-1 models of a number of CAFASP servers.

**MF** (serv) - 81 models for 52 3D targets

## Consensus over transitive PSI-Blast alignments

A. Heger[1], C.A. Wilton[1], and L. Holm[1,2]
*[1] – Institute of Biotechnology, [2] – Department of Genetics, University of Helsinki*
liisa.holm@helsinki.fi

The idea was to use an algorithm for transitive alignment[1], but we kept developing and debugging the server throughout the prediction season. Predictions for targets T0196-T0219 were therefore based on a Blast search against the PDB, predictions for targets T0220-T0252 were based on consensus alignment in the union of the first PSI-Blast[2] neighbour shells of the target and template, and predictions for targets T0253-T0280 can have any number of intermediates between the target and template. No prediction was submitted for a number of the late targets, because the server assumed that the exact target sequence is present in UniProt.

1. Heger,A., Lappe,M. & Holm,L. (2004) Sensitive detection of very sparse sequence motifs. *J. Comp. Biol*., *in press*
2. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.

**mGenTHREADER** (serv) - 320 models for 64 3D targets
**nFOLD** (serv) - 320 models for 64 3D targets

## Fully automated fold recognition using nFOLD and mGenTHREADER

L.J. McGuffin, J.S. Sodhi, K. Bryson & D.T. Jones
*-Bioinformatics Unit, Department of Computer Science, University College London, London WC1H 6BT*
dtj@cs.ucl.ac.uk

There have been a number of improvements in our fully automated fold recognition methods since CASP5. Our popular mGenTHREADER[1,2] method has been improved through the inclusion of profile-profile alignments. We have also developed a new method called nFOLD, that is based on the new mGenTHREADER protocol, but which also incorporates a number of extra inputs into the underlying neural network.

The major change to the original mGenTHREADER algorithm is the implementation of a profile-profile alignment algorithm. The comparison method used was designed to directly compare PSI-BLAST profile scores and is based on an optimized heuristic formula, though essentially comprising a scaled dot product of the two profile vectors. A more minor change is that all alignment parameters (e.g. gap penalties) were optimized using a genetic algorithm to maximize a weighted sum of model quality over a benchmark set of 50 difficult fold recognition targets.

The nFOLD method is an extension of the new mGenTHREADER protocol. Three additional inputs are fed into the neural network which include; the secondary structure element alignment (SSEA) score[2], a new functional site detection score (MetSite)[3] and a simple model quality checking algorithm, MODCHECK[4]. The nFOLD neural network is also trained directly on MaxSub[5] score which allows for a greater assignment of confidence in model quality.

Although the SSEA score has been benchmarked previously as an extra neural network input to mGenTHREADER[2], this is the first time it has been included in a fully automated method within a blind assessment.

The functional site predictions were calculated using a set of classifiers based on the MetSite method[3], which was initially developed in order to predict the location of residues forming commonly occurring metal binding sites in low-resolution structural models. The top ranking MetSite predictions were extracted for the top models generated from the mGenTHREADER profile-profile alignments. Analysis of the MetSite scores showed a significant improvement in distinguishing native and near native-like models from decoy hits and so was therefore implemented as an extra input in the nFOLD method.

The MODCHECK score was also used to directly assess the quality of the models from the profile-profile alignments. The MODCHECK program has been used previously for our CASP predictions[4], however this is the first time it has been implemented in a fully automated method.

A further important improvement to the fold recognition servers has been the implementation of fully automated weekly updates of both the fold recognition library and sequence databases, which reduces the chance that no obvious homologs or fold templates are missed when the PDB is updated.

1. Jones,D.T. (1999) GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* ***287***, 797-815.
2. McGuffin,L.J. & Jones,D.T. (2003) Improvement of the GenTHREADER method for genomic fold recognition. *Bioinformatics*, ***19***, 874-881.
3. Sodhi,J.S., Bryson,K., McGuffin,L.J., Ward,J.J., Wernisch,L. & Jones,D.T. (2004) Predicting metal binding sites in low resolution structural models. *J. Mol. Biol.* ***342***, 307-320.
4. Jones,D.T. & McGuffin,L.J. (2003) Assembling novel protein folds from super-secondary structural fragments. *Proteins: Structure, Function and Genetics* ***53*** (S6), 480-485.
5. Siew,N., Elofsson,A., Rychlewski,L., & Fischer, D. (2000) MaxSub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics.* ***16***, 776-85.

## **MIG_FROST** - 80 models for 29 3D targets

### Toward an efficient threading method

A. Marin[1], J-F. Taly[1], J. Martin[1], R. Andonov[2], S. Balev[3], V. Poirriez[4] and J-F. Gibrat[1]

*1– MIG, INRA, Jouy-en-Josas, 78352, France 2 – Symbiose/IRISA, INRIA,35042 Rennes, France, 3– LIH, U. Le Havre, 76058 Le Havre, France, 4– LAMIH/ROI, U. Valenciennes, 59313 Valenciennes, France*

gibrat@jouy.inra.fr

FROST[3,4] is a fold recognition program based on a sequential use of a series of filters. It consists of 4 components:

- a library of cores representative of all known 3D structures or domains;
- two score functions measuring the fitness of a query sequence for a core;
- a number of algorithms to align the sequence onto the cores;
- a statistical evaluation of the score significance.

Each filter corresponds to a different score function. Since we are using a relatively crude description of the polypeptide chain (each residue is modelled as a single interacting site) it is difficult for a single score function to capture the complex relationship between the amino acid sequence and the 3D structure. Each score function in FROST is supposed to specifically model some particular aspect of this relationship. For the moment, though, only 2 score functions have been fully implemented and tested in FROST.

The first one is based on local parameters. In essence, it is comparable to amino acid substitution matrices, but, because we know the 3D structure of the core, we are able to design matrices that are specific of the residue state in the 3D structure. The state of a residue is defined in terms of secondary structure and surface accessibility to the solvent. With this set of parameters aligning a sequence to a core is akin to align 2 sequences using a set of state-dependent substitution matrices and specific gap penalty (e.g., insertions/deletions are strongly penalized within secondary structure elements).

The second score function uses non-local parameters, i.e., considers interactions between sites in contact in the 3D structure. These parameters are a generalization of the local parameters because we now consider the replacement of a pair of residues in contact in the 3D structure by a pair of residues in the query sequence. The main difficulty with this type of non-local parameters is that one cannot use anymore dynamic programming algorithms to align the sequence onto the core. In fact this alignment problem has been shown to be NP-hard. In the current version of FROST great improvements toward solving, in practice, this problem have been made using linear mixed -integer programming models[1] combined with lagrangian relaxation techniques[2].

The magnitude of alignment scores depends strongly on the sequence length and the 3D features of the cores making them impossible to compare directly. Unlike sequence comparison methods, there is no analytical result available concerning the characteristics of random threading score distributions. To evaluate the significance of the alignment scores we have to calculate empirically such distributions, for each core used in the program, using different query sequence lengths. These distributions permit us to normalize the scores and thus to compare them meaningfully across the complete library of cores. Computing these distributions is very CPU intensive. The availability of fast sequence-structure alignment algorithms is extremely useful in this respect.

Finally, each filter provides a normalized score. A query sequence is thus characterized by a vector of scores. We have to decide, based on this score vector, whether the sequence is compatible with the structure or not. This decision is taken based on a SVM analysis of the results.

1. Andonov,R., Balev,S. and Yanev,N. (2004), Protein threading: From mathematical models to parallel implementation, *INFORMS journal on computing* **16,** 4, Special Issue on Computational Molecular Biology/ Bio-informatics, Greenberg H., Gusfield D., Xu Y., Hart W., Vingro M. Eds.

2. Balev,S. Solving the protein threading problem by lagrangian relaxation, WABI04, 4th Workshop on Algorithms in Bioinformatics, Bergen, Norway, september 14-17, 2004.
3. Marin,A., Pothier,J., Zimmermann,K. and Gibrat,JF. (2002) FROST: a filter-based fold recognition method. *Proteins* **49**, 493-509.
4. Marin,A., Pothier,J., Zimmermann,K. and Gibrat, JF. (2002) Protein threading statistics: An attempt to assess the significance of a fold assignment to a sequence. In *Protein Structure Prediction: Bioinformatics Approach*, (I.F. Tsigelny, ed.), chapter **9**, 227-262, International University Line, La Jolla, CA.

**M.L.G.** - 119 models for 63 3D targets

## Prediction of tertiary structure of proteins based on shadow method

Bo Yang, Ya-dong Wang
*School of Computer Science and Technology, Harbin Institute of Technology, China*
Yangbo@mlg.hit.edu.cn , Yeungbo@gmail.com

This paper reports on a new method, shadow method, for predicting tertiary structure of protein, which introduced the method that people evaluate the object from little information in real life. Our strategy for prediction of tertiary structure of protein is based on the observation that man can guess an answer and testify/overthrow the answer, even to find the most probability answer. We take the second structure of protein as the shadow of tertiary structure, and find the best fitting of shadow to the predicted second structure by other methods.

When man recognizes an object with little information, he always guesses an answer at first. Just like the host guess the guest's identity with his shadow that come from the door left unlocked when a guest come to the door. The host can image a name list that who on this list has a shadow like this one. Moreover he will assume that if someone on the list stand at the door, whether is he/she has the same shadow? Or guess who has the most probability to call in on this time.

For this reason, in first step, we obtained the target's shadow *S*, which is the second structure of target come from the prediction tools, such as PSIPRED, NNPREDICT etc. And we create a name list *A* using PSIPRED which on the list has high structure similarity to the target protein. Then we construct a tertiary structure, prototype *R*, of target referring to *A*. And projected the prototype *R* with DSSP to get its shadow *S'*. Now we have tow shadows, *S* & *S'*. Hereto, the question is 'Are they similar enough?' That's to say 'is our guess reasonable?'

We believe that the more similarity between *S* and *S'*, our guess is more closer to the real identity of the one who after the door.

Of cause, the guess usually fall into fail, we should adjust *R* for a new guess when the difference is distinct between *S* and *S'*. Here, we introduce the evolving algorithm, an optimal algorithm, to adjust *R*.

We design the evolving algorithm as follows:
Step1. Set mutation rate *Pm*; the training's Termination-Conditions: maximum iteration times & the expected precision; Initial Colony *A(N)* with prototype *R*.

Step2. For *i*=1 to N do

Calculate the shadow $S_i^{'}$ of the $i^{th}$ individual in *A*

Step3. Estimate each individual's fitness in *A*, and store the best one of whole to the Elite. Check the Termination-Conditions:
If **True** Jump to Step 7 Else Continue End if

Step4. Using the individuals with high fitness to generate new colony *A'(N)* with Selection operator

Step5. Mutate the individual in *A'* to adjust the each individual's prototype *R*, get the next generation colony *A(N)*

Step6. Repeat the above steps from step2 to step5.

Step7. Return the best result Elite.

In step 3, the fitness indicate that the distance between *S* and *S'*. The smaller distance is, the higher fitness is. And the fitness is evaluated as follows:

$$Fitness(i) = Length(S') - Sum(gap) - |Length(S') - Length(S)|$$

The *Sum(gap)* denote that the number of gap in *S'*.

In step 5, the Mutation take place on the points with low score of stability. The score indicate that the sameness of the point between *S* and *S'*. Those points of prototype with low score should be adjusted to mutate a better prototype to fit the *S*.

At last, we can get a guessed prototype *Elite* when the evolving accomplished. We take regard the *Elite* as the most probability prototype of the guest after the door.

The future work:
In our algorithm didn't use the energy minimization to optimal the final result, so we expect to get better result with the energy minimization in future. And we will extend this method to be a Server for predicting of tertiary structure of protein based on our secondary structure prediction Web Service at http://mlg.hit.edu.cn/xml


1. Qian,N., Sejnowski,T.J. (1988). Predicting the secondary structure of globular proteins using neural network models, *J. Mol. Biol.* **202**, 865-884.
2. Whitley,D. (1989) The GENITOR Algorithm and Selection Pressure: Why Rank-Based Allocation Reproduction Trials is Best, In: Schaffer, J. (Editor), Proceeding of the 3rd International Conference on Genetic Algorithm, Morgan Kaufmann Publishers, Los Altos, CA.
3. Gibas,C., Jambeck,P. (2002). Developing bioinfomatics computer skills. Jointly published by O'Reilly & Associates, Inc. and Science Press.
4. Baldi,P., Brunak,S. (2003). Bioinformatics: The Machine Learning Approach, published by arrangement with MIT through Arts & Licensing International, Inc., USA.
5. Bo,Y., Yadong,W., Xiaohong,S., Lijuan,W. (2004). Solving Flat-Spot Problem in Back-Propagation Learning Algorithm based on Magnified Error. Proceedings of the Third International Conference on Machine Learning and Cybernetics, Shanghai, China, 26-29 August, 2004, vol.3, 1784-1788.


## **MPM** - 25 models for 25 3D targets

### Comparative modeling of CASP6 target proteins


J. Kopp, J.N.D. Battey, L. Bordoli and T. Schwede
*Biozentrum Basel and Swiss Institute of Bioinformatics, University Basel, Switzerland*
Torsten.Schwede@unibas.ch


We aimed at building comparative models for CASP6 targets where templates could be identified for at least part of the target sequence. Since template selection and target-template alignment are considered as the crucial steps in comparative modeling, we used a "build many - select best" strategy: several methods for template selection, alignment and model building were applied in parallel to generate an ensemble of models. These were evaluated to identify the best candidate for subsequent rounds of iterative model improvement. Models found to be contradictory with available biological information (e.g. incomplete metal binding sites) were not submitted.

*Template Selection*: Templates were selected from the SWISS-MODEL template library[1] using sequence based search methods: First, templates sharing high sequence similarity were identified using PSI-BLAST[2] with a target sequence profile based on NR. Target sequence regions for which no template was identified in the previous step were used to generate a target Hidden Markov Model using SAM 3.4[3,4] for searching the template library.

*Target-Template Alignment:* Multiple sequence alignments for the target-template sequence family were generated using the following three methods: a) T_COFFEE including information from structural alignments of related templates[5], b) a template sequence HMM generated by SAM[3,4], and c) profiles for both target and template generated with SAM were aligned with LOBSTER[6] or COMPASS[7].

*Model Building and Evaluation*: Models for the resulting alignments were built based on single templates using both SWISS-MODEL[1] in project mode and Modeller [8]. Following a "build many - select best" strategy, the best model for subsequent rounds of iterative model improvement was selected by evaluation with the atomic mean force potential ANOLEA[9], as well as Gromos96 force field energy after steepest descent minimization[10].

*Model Validation and Iterative Refinement*: Ranking and selection of possible template structures, and the target-template alignment of the best-scoring model was cross-validated with PFAM[11] and TIGRFAMs[12] profiles, and other available biological information (e.g. motivation for modeling T0240 as monomer). Regions identified as unreliable during the evaluation steps were subjected to a refinement process: Alignment modification, loop re-modeling, and re-arrangement of side-chain conformations were applied iteratively until the ANOLEA evaluation converged.


1. Schwede,T., Kopp,J., Guex,N. & Peitsch,M.C. (2003) SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res.* **31**, 3381-3385.
2. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.

3. Hughey,R. & Krogh,A. (1996) Hidden Markov models for sequence analysis: Extension and analysis of the basic method, *CABIOS* **12**, 95-107.
4. Karplus,K., Barrett,C. & Hughey,R. (1998). Hidden Markov Models for Detecting Remote Protein Homologies, *Bioinformatics* **14**, 846-856.
5. Notredame,C., Higgins,D. & Heringa,J. (2000). T-Coffee: A novel method for multiple sequence alignments. *J. Mol. Biol.* **302**, 205-217.
6. Edgar,R.C. & Sjolander,K. (2004) COACH: profile-profile alignment of protein families using hidden Markov models. *Bioinformatics* **20**, 1309-1318.
7. Sadreyev,R. & Grishin,N. (2003) COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J. Mol. Biol.* **326**, 317-336.
8. Sali,A. & Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779-815.
9. Melo,F. & Feytmans,E. (1998) Assessing Protein Structures with a Non-local Atomic Interaction Energy. *J. Mol. Biol.* **277**, 1141-1152.
10. Van Gunsteren,W. (1996) Biomolecular Simulations: The GROMOS96 Manual and User Guide. VdF Hochschulverlag ETHZ.
11. Bateman, A. et. al (2004) The Pfam Protein Families Database, *Nucleic Acids Res.* **32**, D138 - D141.
12. Haft,D.H., Selengut,J.D. & White,O. (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res.* **31**, 371-373.

## **MUMSSP** - 9 models for 2 3D targets

### How do the web facilities help predictors from head to toe of homology modeling?

M.R. Saberi, A. Razzazan, H. Ramezani and A. Baratian
*Medicinal Chemistry Division, School of Pharmacy, Mashhad University of Medical Sciences, Mashhad, Po. Box: 91775-1365, Iran*
saberimr@mums.ac.ir

In this project, we applied the theory of evolution method, including threading and comparative modeling. It was carried out through NCBI[1; 2], Swiss-Prot[3], EMBL[4], PDB[5], SCOP[6], CATH[7] and a dozen of related web sites that perform single and multiple alignments to get similar sequences and find proper template(s) as well as other tasks in bioinformatics field.

Similarity search was carried out through PSI-BLAST[8] and PHI-BLAST[8] against nr and PDB to find high identical proteins as the first line similarity and homology study as well as finding proper templates based on sequence-sequence alignment. These methods were applied mainly through ExPASy[3], NCBI[2] and EBI[9] services. A rang of different PAM and BLOSUM thresholds were applied as similarity search matrices. Some computer based programs such as ClustalX[10], ViewerLite, MODELLER[11] and SPDBViewer[12; 13] were applied to produce and analyze sequence alignments in both multiple and single routes to find conserved and identical regions within the query and similar sequences. Different gap penalties were exploited to improve alignments when needed. Alignments were deeply studied to find critical segments which might play a key role in the functionality of the proteins. In the next step, we predicted the possible secondary structure for the query sequences. This was carried out through Jpred[14], 3D-PSSM[15] and PSIpred[16]. Resolution and R-factor of a crystallographic structure were indicative of the accuracy of the structure. Templates were carefully considered regarding their folding and family in SCOP and CATH servers. Threading method came into account when proper template(s) did not come across from PDB-BLAST. This was employed through FUGUE[17] and 3D-PSSM servers. FUGUE program, scan a database of structural profiles, calculate sequence-structure compatibility scores and produce a list of potential homologues proteins and alignments.

Having predicted conserved areas of the query secondary structure and proper templates in hand, models were created in MODELLER 6v2 on a high performance PC platform by satisfaction of spatial restraints. Hundreds of models were generated using almost all scripts of MODELLER such as FULL_HOMOL, MULTIPLE_MODELS, SEGMENT_MATCHING, MAKE_RESTRAINTS and REFINE. High speed internet connection let us to evaluate the models on web based evaluation programs such as ERRAT[18], VERIFY3D[19], WHAT_CHECK[20], WHAT IF[21] and iMOLTALK[22] on UCLA, BIOTECH and ExPASy servers. Models were investigated in SPDBViewer and ViewerLite programs checking amino acids making clash, Phi-Psi angles, secondary structure matching the secondary structure prediction etc. before submission to evaluation sites. Although the group took advantage of some commercial packages such as MOE[23] but we preferred to use downloadable programs to prove the power of pure web based bioinformatics in homology modeling. The said programs allocated atom environment, solvent accessibility and stereochemistry of models. Models were modified in MODELLER when needed and the last steps were repeated to improve the protein structure. Models from CPHmodels[24], ESyPred3D[25] and SWISS-MODEL[26] were compared to our models to refine and confirm the folding and improve the models. The accuracy of the various models from different methods was

relatively similar. Other factors such as template selection and alignment accuracy usually showed a larger impact on the model accuracy.


1. Jenuth,J.P. (2000). The NCBI. Publicly available tools and resources on the Web. *Methods in Molecular Biology* **132**, 301-12.
2. Wheeler,D.L., Church,D.M., Federhen,S., Lash,A.E., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E., Tatusova,T.A. & Wagner,L. (2003). Database resources of the National Center for Biotechnology. *Nucleic Acids Res.* **31**, 28-33.
3. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I., Pilbout,S. & Schneider,M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365-70.
4. Kulikova,T., Aldebert,P., Althorpe,N., Baker,W., Bates,K., Browne,P., van den Broek,A., Cochrane,G., Duggan,K., Eberhardt,R., Faruque,N., Garcia-Pastor,M., Harte,N., Kanz,C., Leinonen,R., Lin,Q., Lombard,V., Lopez,R., Mancuso,R., McHale,M., Nardone,F., Silventoinen,V., Stoehr,P., Stoesser,G., Tuli,M.A., Tzouvara,K., Vaughan,R., Wu,D., Zhu,W. & Apweiler,R. (2004). The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.* **32 Database issue**, D27-30.
5. Sussman,J.L., Lin,D., Jiang,J., Manning,N.O., Prilusky,J., Ritter,O. & Abola,E.E. (1998). Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. *Acta Crystallogr D Biol Crystallogr* **54**, 1078-84.
6. Ahn,G.T., Kim,J.H., Hwang,E.Y., Lee,M.J. & Han,I.S. (2004). SCOPExplorer: a tool for browsing and analyzing structural classification of proteins (SCOP) data. *Molecular Cell* **17**, 360-4.
7. Pearl,F.M., Bennett,C.F., Bray,J.E., Harrison,A.P., Martin,N., Shepherd,A., Sillitoe,I., Thornton,J. & Orengo,C.A. (2003). The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic Acids Res.* **31**, 452-5.
8. Jones,D.T. & Swindells,M.B. (2002). Getting the most from PSI-BLAST. *Trends in Biochemical Sciences* **27**, 161-4.
9. Rodriguez-Tome,P. (2001). EBI databases and services. *Molecular Biotechnology* **18**, 199-212.
10. Li,K.B. (2003). ClustalW-MPI: ClustalW analysis using distributed and parallel computing. *Bioinformatics* **19**, 1585-6.
11. Sanchez,R. & Sali,A. (2000). Comparative protein structure modeling. Introduction and practical examples with modeller. *Methods in Molecular Biology* **143**, 97-129.
12. Guex,N. & Peitsch,M.C. (1997). SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* **18**, 2714-23.
13. Kaplan,W. & Littlejohn,T.G. (2001). Swiss-PDB Viewer (Deep View). *Brief Bioinformatics* **2**, 195-7.
14. Cuff,J.A., Clamp,M.E., Siddiqui,A.S., Finlay,M. & Barton,G.J. (1998). JPred: a consensus secondary structure prediction server. *Bioinformatics* **14**, 892-3.
15. Bates,P.A., Kelley,L.A., MacCallum,R.M. & Sternberg,M.J. (2001). Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. *Proteins* **Suppl 5**, 39-46.
16. McGuffin,L.J., Bryson,K. & Jones,D.T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics* **16**, 404-5.
17. Shi,J., Blundell,T.L. & Mizuguchi,K. (2001). FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol.* **310**, 243-57.
18. Colovos,C. & Yeates,T.O. (1993). Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Sci.* **2**, 1511-9.
19. Luthy,R., Bowie,J.U. & Eisenberg,D. (1992). Assessment of protein models with three-dimensional profiles. *Nature* **356**, 83-5.
20. Hooft,R.W., Vriend,G., Sander,C. & Abola,E.E. (1996). Errors in protein structures. *Nature* **381**, 272.
21. Vriend,G. (1990). WHAT IF: a molecular modelling and drug design program. *Journal of Molecular Graphics* **8**, 52-56.
22. Diemand,A.V. & Scheib,H. (2004). iMolTalk: an interactive, internet-based protein structure analysis server. *Nucleic Acids Res.* **32**, W512-6.
23. Bajorath,J. (2001). Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *Journal of Chemical Information and Computer Sciences* **41**, 233-45.
24. Lund,O., Nielsen,M., Lundegaard,C. & Worning,P. (2002). *Abstract at the CASP5 conference A102*.
25. Lambert,C., Leonard,N., De Bolle,X. & Depiereux,E. (2002). ESyPred3D: Prediction of proteins 3D structures. *Bioinformatics* **18**, 1250-6.
26. Schwede,T., Kopp,J., Guex,N. & Peitsch,M.C. (2003). SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Research* **31**, 3381-5.

## **MZ_2004** - 64 models for 64 3D targets

### Energy based 3D protein structure predictions

Koji Ogata[1], Raphael Leplae[2] and Shoshana J. Wodak[2,3]
*[1] - Zoegene Corp., Japan; [2] - Service de Conformation de Macromolecule Biologique et Bioinformatique, Université Libre de Bruxelles, Belgium;*
*3 - University of Toronto/The Hospital for Sick Children, Canada*
mz@scmbb.ulb.ac.be

ModzingerZ (MZ) is a software package dedicated to the prediction of protein structures by homology modelling. Structural templates are identified by a two steps procedure. A first set of structural template candidates for the target sequence are identified using Psi-BLAST[1] with default parameters and 5 times iterations against a sequence database combining sequences from GenBank[2] and PDB-sub (PDB-sub containing sequences with <90% sequence identity from PDB). In the second step, individual PDB entries obtained from the first step are used as query sequence against PDB-sub with Blast to identify additional homologs with known 3D structure. All the identified template candidates are then structurally aligned. A profile is derived from the structural alignments and the target sequence is aligned against this profile[3]. In addition a sequence profile is computed for each identified structural template by running Psi-Blast against the GenBank sequence database and pruning so as to leave highly similar sequences (with identity more than 50% and less than 100%). In performing these alignments, gaps inside the secondary structure elements (computed using DSSP[4]) were penalised.

Structurally conserved regions (SCR) in the target sequence were then defined as residues aligned to those of the structural templates displaying an RMSD ≤1.0Å in the corresponding multiple structural alignment. The backbone of these SCR residues in the target sequence was built using the main chain coordinates of the template with the highest BLOSUM62 score to the target. Side chain coordinates from the same template were also used whenever the amino acid of the target and template were the same.
The remaining regions in the target sequence, called structurally variable regions (SVR), were built by using the main chain atom coordinates of the template structure having the highest BLOSUM62 score computed without insertion/deletion regions. For regions with insertions/deletions, an energy-based loop modelling method[5] was used to find suitable loop conformations. The force-field, used for evaluating the conformations, models each residue by two interaction centers positioned at the Ca and Cb atoms. The pairwise interaction energies between these centres was derived by computing the average of the potential energy of the AMBER force field[6] for main chains and side chains interactions for all residue pairs found in the PDB. We verified that this force-field yields rather accurate predictions for individual protein loops as well as several interacting loops. This loop modelling approach can be applied to segments of maximum 22 residues; longer loops were simply not modelled.

Residues without side chain coordinates from a template structure were generated using the Monte Carlo method with the AMBER force field.

Models produced by the above procedure were examined, and the alignment was adjusted (either manually or with alignment tools), whenever some inconsistencies (on the sequence, structure or biological level) were discovered. The new alignment was then re-fed to the model building method described above.

1. Altschul,S.F. et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25** *(17)*, 3389-3402.
2. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A., Wheeler,D.L. (2002) GenBank. *Nucleic. Acids. Res.* **30**, *17-20*
3. Rychlewski,L., Jaroszewski,L., Li,W., Godzik,A. (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci.* **9**(2), *232-41*.
4. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22** *2577-2637*.
5. Ogata,K., Leplae,R., Wodak,S.J. An Energy Based Predictions for Multi-loops of Proteins. *in preparation*.
6. Weiner,S.J., Kollman,P.A., Nguyen,D.T. and Case,D.A. (1986). An all atom force field for simulations of proteins and nucleic acids. *J Comput Chem.* **7**, *230-252*.

## **NIM_CASP6** - 13 models for 13 3D targets

### Functional network analysis as an effective scoring system for protein structure prediction

C. Morales-Almonte[1] and G. del Rio[1]
*Instituto de Fisiologia Celular/UNAM, Mexico*
gdelrio@ifc.unam.mx

Automatic prediction of protein structure requires the evaluation of multiple models. As a consequence, reliable scoring systems to identify native-like protein structures from these models are essential for achieving accurate predictions. We propose that a scoring system that identifies unique characteristics to every protein may be more reliable that the current scoring systems used for protein structure prediction based on average characteristics of protein structure. A proof for this idea has been previously presented by Valencia and col.[1]. However, such approach was only successful for short proteins (<170 amino acids). Here we describe an alternative approach to Valencia's that is not dependent on the protein length.

Our approach, dubbed NIM, is based on the assumption that every protein has a unique set of critical residues for the protein's function[5]. Critical residues may be identified from protein sequences using phylogenetic approaches, while we have described a highly specific method to identify critical residues from protein structures[2]. Our scoring system then, determines the quality of a protein structure model by matching the critical residues observed in the model with those determined from the protein sequence by phylogenetic approaches. To identify the critical residues from protein structures, we represent the structures as a network of residue contacts at 5 or less Angstroms. From this representation, we identify the most traversed residues in the network by counting the number of times a residues is transited in connecting every pair of residues in the network through the shortest path, using Dijkstra's algorithm. We have found that the most traversed residues match with the critical residues for protein function[2].

To evaluate the reliability of NIM, we participated in CASP5 and CASP6. In CASP5 we compared our method with a scoring system based on an energy function, PROSPECT2[3]. For CASP6 we are now comparing our method to BLASTPGP[4]. BLASTPGP aligns a protein target with every protein of known structure (template) and scores these based on the observed sequence identity of the alignment. We learned in CASP5, that our scoring system improved the predictions reported by PROSPECT2. A similar trend was observed in CASP5 and CASP6: Protein targets presenting high sequence similarity to a protein template, NIM, PROSPECT and BLASTPGP predicted the same fold, but differed as the similarity felt down. We are developing a server to give access to the scientific community to our scoring system.

In summary, we have developed a new scoring system for protein structure prediction. Our approach may represents a new kind of scoring system that has shown to be useful in improving some of the current methods for fold recognition.


1. Olmea,O, Rost,B, Valencia,A. (1999). Effective use of sequence correlation and conservation in fold recognition. *J. Mol. Biol.* **293**, 1221-39.
2. del Rio,G., del Rio,H., Bartley,T., Castro-Obregon,S., Bredesen,D.E. Functional assessment of protein structures as biological networks. *Submitted to FORCASP*.
3. Kim,D., Xu,D., Guo,J., Ellrott,K., Xu,Y. (2003) Prospect II: Protein Structure Prediction Program for the Genome-Scale Application. *Protein Engineering* **16**, 641-650.
4. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
5. del Rio,G., Garciarrubio,A., Cusack,M, Bredesen,D.E. (2004) Functional network analysis as an effective scoring system for protein structure prediction. *Submitted to FORCASP*.


## **Oka** - 125 models for 62 DP / 63 DR targets

### Prediction of domain boundaries and disordered regions in proteins with unknown tertiary structure


S.O. Garbuzynskiy, M.Yu. Lobanov, D.N. Ivankov,
N.S. Bogatyreva, A.V. Finkelstein and O.V. Galzitskaya
*Institute of Protein Research RAS*
ogalzit@vega.protres.ru


Our method of prediction of domain boundaries and disordered regions in query proteins is based on calculating profiles (using our program PROFILE) where one of 20 numbers is attributed to each residue according to its type.

Domain boundaries were predicted as follows. We formed a database of multidomain proteins (proteins with at least one domain boundary) with sequence identity below 25% taking them from the SCOP[1] database. Positions of domain boundaries were also obtained from SCOP. Then we calculated the occurrence of each of 20 types of amino acid residues at the domain boundaries as compared to the occurrence in all proteins of our database. Using the obtained 20 numbers we calculated the profile for a query protein. One of the 20 numbers was assigned to each residue of the query protein; then, these numbers for the residues inside the window of 41 residues were averaged and the mean number was attributed to the central residue of the window. Thus we have a profile where maxima should correspond to the domain boundaries.

An alternative scale for domain predictions[2] was produced using an approach based on the assumption that the unique tertiary structure of protein is a result of the balance between the gain of native interactions and the loss of conformational entropy of the unfolded chain. In other words, the topology of the chain determines how much entropy is lost while native interactions are formed. So it can be suggested that high side chain entropy of a region in a protein chain should be compensated by high interaction energy within the region, which could correlate with a well-structured part of the globule, that is, with a domain unit. This means that domain boundaries are composed of mainly amino acid residues with low conformational entropy. Considering the conformational entropy as the number of degrees of freedom on the $\phi$, $\psi$, and $\chi$ angles for each amino acid along the chain, our method for domain boundary prediction relies on finding the minima in a latent entropy profile.

Possible information about homologs of query proteins was also used in our predictions. If a close homolog of a query protein had a known 3D structure, we took into account the available information about the domain boundaries in that homolog. If no 3D structures of homologs are available, we sometimes constructed multiple alignments using PSI-BLAST[3] searching for possible evolutionary units in query proteins. We consider an evolutionary unit as a part of protein which is observed either in isolation or as a part of different multidomain proteins. Since it is one of the definitions of a domain[1], the presence of more than one evolutionary unit in a target protein may indicate that it is probably a multidomain protein. If, for example, only the first part of a query sequence is aligned with one group of proteins while only the second part is aligned with another group, it is evidence that the query protein is a two-domain one.

For making our prediction of the disordered regions in target proteins the profile was constructed using a scale of an expected number of contacts[4] for each of 20 types of residues in a globular state. The idea is that amino acid residues, forming disordered regions of proteins, undoubtedly make fewer contacts per residue than residues in ordered regions in the native globular state. It is obvious that residues of different types usually form an unequal number of contacts (Trp generally makes more contacts than Gly). So we can try to predict the number of contacts per residue starting from sequence only. The scale of the number of contacts for 20 types of amino acid residues in globular state was constructed as follows. We selected a database of protein domains with less than 80% sequence identity values using SCOP. Then the average number of residue-residue contacts per residue of each of 20 types was calculated with an assumption that two residues are in contact if any pair of their heavy atoms (i.e. at least one atom per residue) is situated at a distance less than 8.0 Å from each other. The scale obtained in such a way was used for constructing the profile of a query protein; the regions on the profile with a low estimated number of contacts per residue were predicted as possible unstructured regions.


1. Lo Conte,L., Brenner,S.E., Hubbard,T.J.P., Chotia,C. & Murzin,A.G. (2002). SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.* **30**, 264-267.
2. Galzitskaya,O.V. & Melnik,B.S. (2003). Prediction of protein domain boundaries from sequence alone. *Protein Sci.* **12**, 696-701.
3. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
4. Garbuzynskiy,S.O., Lobanov,M.Yu. & Galzitskaya,O.V. (2004). To be folded or to be unfolded? *Protein Sci.* **13**, in press.


## **Pan** - 272 models for 64 3D targets

### Using secondary structure to build structural template


Y. He[2], S. Qin[2], X.M. Pan[1,2,*], M. Beckstette[3] and R. Giegerich[3]

*[1]-Department of Biological sciences and biotechnology, Tsinghua University, Beijing, China, [2]-National Laboratory of Biomacromolecules Institute of Biophysics, Chinese Academy of Sciences, Beijing, China, [3]-AG Praktische Informatik Technische Fakultät, Universität Bielefeld, 33594 Biefeld, Germany*

xmpan@sun5.ibp.ac.cn


Homology modeling is an effect method for structure prediction when suitable template protein exists; but sometimes, PSI-BLAST[1] can not find proper homologous because of the low sequence identity or bad structural quality and in such cases homology modeling is always impossible. Herein we describe a method for detecting distant homologous which have low sequence identity with the target protein but may share the same fold patterns by involving the structure information into the sequence alignment.

We use the multiple linear regression (MLR) method to predict secondary structure from the amino acid sequence that was reported previously[2]. For the recent months, the implementation of this prediction method has been changed a lot; the new implementation adopts the PSSM (Position Specific Scoring Matrix) generated by PSI-BLAST as its only input information. In addition, the

``Jury system" adopted in the old implementation was obsolesced by the new implementation, with a new engine replaced it. The new implementation has achieved an average accuracy better than 80% in the prediction for a set of about 1400 protein chains (unpublished results).

We also use MLR method to predict relative solvent accessibility, and the implementation of this method was reported previously[3], and we now have developed a new implementation of it which has achieved an average accuracy better than 80% in the prediction for a set of about 1116 protein chains at a threshold of 20% for the definition of two-state of solvent accessibility (unpublished results).

A protein referred as a homologue, not only for its homology of amino acid sequence, but also more conservation at the structural level. There are two strategies to find suitable templates for homology modeling by involving the structure information.

One strategy in this study is still based on the sequence-driven detection, searching the target sequence against the sequence database compiled from a representative PDB collection. A reduced alphabet is employed which divides the twenty types of amino acid into eight groups[4]. This reduced alphabet can increase the possibility of detection for distant-homology protein; meanwhile, it can increase the possibility of false positives. A restriction condition of high similarity of secondary structure of the target and that of potential homologous is employed to exclude the false positives. A 6x6 score matrix is introduced into the alignment procedure of the secondary structure, each class of all three states of secondary structure (H, E, C) is divided into two types according to states of two adjacent residues: at edge, or not at edge. Since the prediction is relatively weak for those residues at edge, the assignment of secondary structure states for residues at edge and those not at edge should be treated differently.

Another strategy is based on the structure-driven pattern detection, searching the secondary structure pattern of the target against that secondary structure library compiled from the same collection of chains. Segments predicted as coils are not very confidential, so their properties of solvent accessibility are surveyed. We combine predicted secondary structure and solvent accessibility as well as sequence into a score matrix for search and alignment. The score matrix employed is an expanded matrix from the 6x6 score matrix above. The 6x6 matrix has 36 blocks, and each block will be further divided into 20x20 blocks with values derived from BLOSUM62 matrix, the solvent accessibility is also included, for exposed, buried and uncertain state which is a critical state between exposed and buried, different scores are appended respectively, finally it is a 120x120x3 matrix.

Both of the two searching strategies will produce alignments of the target versus the template, and can be directly used by MODELLER[6] to build models.


1. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
2. Pan,X.M. (2001). Multiple linear regression for protein secondary structure prediction. *Proteins.* **43**(3), 256–259.
3. Li,X. & Pan,X.M. (2001). New method for accurate prediction of solvent accessibility from protein sequence. *Proteins*. **42**(1), 1–5.
4. Pan,X.M., Niu,W.D. & Wang,Z.X. (1999). What is the minimum number of residues to determine the secondary structural state? *J Protein Chem.* **18**(5), 579–84.
5. Šali,A. & Blundell,T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*. **234**, 779–815.


## **Panther** - 55 models for 28 3D targets

### Backbone clusters as structural templates


Hao Wang, Robert W. Harrison

*Department of Computer Science, Georgia State University*


One recurring critical problem revealed in CASP has been the ability to model insertions and deletions in protein structure. Related to this is the inability of potential based modeling approaches to correct for minor sequence alignment errors. Two approaches were tested to see if they had potential to help overcome these issues. The first approach was to extend the molecular mechanics potential by including a mean-force potential. The potential was chosen by defining a set of most common nodal or “eigenstructures” together with terms to represent the range of variation in the structure. These nodal structures effectively span the space of allowed and observed peptide conformations. The problem of modeling an insertion or deletion then becomes the problem of identifying the correct nodal structure. The nodal structures were chosen via K-nearest neighbors clustering to provide a uniform covering of the space of structures. The second approach was to add a switching hydrogen bond potential to help stabilize the backbone structure. This potential was implemented with a Morse function.

Clustering of Protein Backbone Structure

The protein database was analyzed by K nearest neighbors (Knn) clustering on α–carbon atoms. The distances between all pairs of amino acids within short fragments of the structure were used as a basis for clustering. The distance matrices were generated and clustered. For 5-mers 100 clusters were enough to completely cover the space of conformations, and for 10-mers between 100 and 1000 were sufficient. With 1000 clusters and 10-mers the clusters index the space of protein structures to an accuracy of 0.4454Å. The clusters on α-carbon atoms were then used as a basis to extract and index distances between other atoms in the protein backbone. Experimental trial showed that O-N, C-O, C-N and O-O distances were sufficient to build the protein backbone with good local geometry from α-carbon positions.

| Window Size | Number of Clusters | Root Mean Square error | Root Mean Square Error (Chiral Cluster) |
|---|---|---|---|
| 5 | 10 | 0.2469 | 0.3510 |
| 5 | 100 | 0.1425 | 0.2101 |
| 5 | 1000 | 0.0983 | n.d. |
| 10 | 10 | 0.8606 | n.d. |
| 10 | 100 | 0.5666 | 0.7050 |
| 10 | 1000 | 0.4454 | 0.5566 |
| 10 | 10000 | 0.3087 | n.d. |

There is a problem in the use of clustering based on distances alone. The distance matrix is achiral, and therefore the clusters may reflect a mixture of structures. Post-CASP calculations using a chiral cluster, where chirality was implemented with a scaled triple product or pyramid height term, show an increase in the RMSE at the same number of clusters. This suggests that the achiral clusters are partial mixtures of structures and including the chirality will improve the accuracy of the approach. However it also shows that more clusters will be needed to achieve increased accuracy.

Distance Restraints from α-Carbon Clusters

The clusters are applied by finding the closest cluster based on the distances between all observed pairs of α-carbon atoms in the starting model structure. Distances from unobserved atoms are ignored in this calculation. Typically the average difference between the closest cluster and the model is about 0.5Å or less. All overlapping fragments are used to determine the distances, and they typically define a range of values that are possible for a given fragment. The distance constraints are implemented with the split-harmonic potential in AMMP. This term was originally introduced into AMMP for representing NOE-based distance restraints, and to support solving NMR structures.

Using the Distance Restraints

The distance restraints are applied throughout the modeling building steps in AMMP. Initially the new parts of the model, which correspond to atoms not present in the starting model including side chain and amino acid insertions, are built in the context of a static known structure. After building and energy minimizing the new parts of the structure, the entire model is allowed to move.

Hydrogen Bonding via Morse Potentials.

Hydrogen bonds stabilize regular protein structures like helices and sheets. One of the best simple visual checks on the correctness of a model is whether the regular secondary structure is conserved. While it does not prove correctness, it is highly diagnostic of procedural errors when regular structure is disrupted in the modeling process. Therefore, we hypothesized that reinforcing regular structures by increasing the hydrogen bonding terms above the default values in our molecular mechanics force field would improve the quality of the models. In order to avoid disrupting the structure by simply increasing charges or changing the Van derWaals terms, a bonding potential that disassociates was chosen. The simplest such potential is the Morse potential <insert equation here>. The radius was 2.4Å, the potential depth to 2 kcal/mol, and the bond order was set to 1. These values were found by adjustment to preserve structure when energy minimizing a protein structure. Morse bonds were defined for all pairs of backbone hydrogen bonds in the protein.

## **Preissner_Steinke** - 123 models for 60 3D / 7 FN targets

### A distributed pipeline for structure prediction

E. Michalsky [1,2], A. Goede [1,2], R. Preissner[1,2], P. May[1,3] and T. Steinke[1,3]

[1] - *Berlin Center for Genome Based Bioinformatics*, [2] - *Charité, University Medicine Berlin, Germany*, [3] - *Zuse-Institut Berlin (ZIB), Germany*

elke.michalsky@charite.de

The first step in our protein structure prediction procedure is to identify suitable templates for homology modeling. A pipeline was established to perform successive PSI-Blast[1] searches automatically in order to find template structures. If no suitable template structure was found in the Protein Data Bank[2] (PDB), a PSI-Blast search in SwissProt[3-4] was performed to initiate a further Blast search in the PDB starting from the SwissProt hits. Here, it was tried to collect several good Blast hits having the same PFAM domain in order to be able to construct multiple alignments from them. If the Blast search in

SwissProt had found several proteins with the same (known) function, a new search among the PDB structures was initiated to find protein structures having the same function. Moreover, we collected secondary structure predictions from different resources and used them to choose suitable and to eliminate implausible templates from the list of Blast hits. Also the fold prediction provided by JCSG (Joint Center for Structural Genomics) via the CASP6 homepage was incorporated into the template search.

Starting with the templates found with the Blast searches, the Blast Alignments were refined manually, focusing on the conservation of secondary structures, i.e. gaps within secondary structures were avoided. Here again, the secondary structure predictions were incorporated. If PDB structures with bound ligands were available, the amino acid residues responsible for the binding, and thus for the function of the protein, were identified and the alignment was inspected towards conservation of those residues. Function predictions were derived using this information and with aid of the Columba database of protein structure annotation[5].

To obtain reasonable alignments using entire available protein family information, we used STRAP, which is a tool for generating multiple structure based alignments, developed in our research group at Charité[6]. Gaps, i.e. insertions as well as deletions in the alignment, were handled with the tool LIP (**L**oops **I**n **P**roteins)[7]. The program LIP is based on a comprehensive compilation of backbone conformations from a recent version of the PDB. In the first step protein segments are selected that fit approximately into the gap in the protein structure and that have the required number of amino acids. In order to evaluate the fitting, for each segment a goodness is calculated. The goodness is defined as the RMSD between a loop candidate and the gap in the protein structure with respect to the distance between the stem residues and several certain dihedral angles. Thereafter, the selected protein segments are evaluated using an optimized scoring function. Besides the goodness, it includes additional values, i.e. the RMSD between the stem residues as well as a sequence alignment score based on a modified BLOSUM mutation matrix. Clashes of the new loop with the core of the protein are avoided. The best-ranked protein segment is inserted into the gap between adjacent secondary structures.

After filling the gaps in the protein models, mutations, side chain rotamer selection and successive energy minimizations were performed by means of the protein visualization and modeling tool Swiss-PdbViewer, version 3.7b[8]. Remaining protein segements for which no suitable template had been found, were predicted using special Blast searches for short nearly identical segments and with aid of the secondary structure predictions.

If in the pipeline described above no suitable template structure was found, a protein threading procedure using the Theseus[9] implementation was initiated at ZIB. The target sequence was scanned for potential multi-domain proteins using Domain-Fishing[10].

Theseus is a parallel implementation of a protein threading based on a branch-and-bound search algorithm to find the optimal threading through a library of template structures. The template fold library is built on SCOP[11] domains, which are available as ASTRAL[12] PDB-style files. Theseus uses a template core model based on secondary structure definition and a scoring function based on pseudo energies that include pairwise contacts, solvent accessibility, homology, variable gap lengths, and secondary structure matching between template and target as predicted by PsiPred[13-14]. From the highest scoring templates we selected the most probable template for further processing.

The reconstructed loops were modeled with the LIP tool[7]. Side chain rotamers were (partly) selected using Swiss-PdbViewer or SYBYL/Biopolymer[15]. The obtained initial structural guess was refined by a local optimization protocol and a final short energy minimization using the Tripos60 force-field and AMBER charges as implemented in SYBYL/Biopolymer.


1. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W., Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389-3402.
2. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N., Bourne,P.E. (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235-242.
3. Bairoch,A., Apweiler,R. (1997). The SWISS-PROT protein sequence database: its relevance to human molecular medical research. *J. Mol. Med.* **75**, 312-316.
4. Bairoch,A., Boeckmann,B., Ferro,S., Gasteiger,E. (2004). Swiss-Prot: juggling between evolution and stability. *Brief Bioinform.* **5**, 39-55.
5. Rother,K., Mueller,H., Trissl,S., Koch,I., Steinke,T., Preissner,R., Froemmel,C., Leser,U. (2004). COLUMBA: Multidimensional data integration of protein annotatios. DILS conference on databases in life sciences, LNBI 2994, 156-171.
6. Gille,C., Lorenzen,S., Michalsky,E., Frommel,C. (2003). KISS for STRAP: user extensions for a protein alignment editor. *Bioinformatics* **19**, 2489-2491.
7. Michalsky,E., Goede,A., Preissner,R. (2003). Loops in Proteins (LIP) - a comprehensive database for homology modeling. *Prot. Eng.* **16**, 979-985.

*8.* Guex,N., Peitsch,M.C. (1997). SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis* **18**, 2714-2723.
*9.* May,P., Steinke,T., Meyer,M. (2004) THESEUS: A Parallel Threading Core. *Proceedings of the 12th Internat. Conf. on Intelligent Systems for Mol. Bio. (ISMB)* and t*he 3rd European Conf. on Comp. Bio. (ECCB)* Poster K55,199.
*10.* Contreras-Moreira,B., Bates,P.A. (2002). Domain Fishing: a first step in protein comparative modelling. *Bioinformatics* **18**, 1141-1142.
*11.* Lo Conte,L., Brenner,S.E., Hubbard,T.J.P., Chothia,C., Murzin,A. (2002). SCOP database in 2002: refinements accommodate structural genomics. *Nucl. Acids Res.* **30**, 264-267.
*12.* Chandonia,J.M., Walker,N.S., Lo Conte,L., Koehl,P., Levitt,M., Brenner,S.E. (2002). ASTRAL compendium enhancements. *Nucl. Acids Res.* **30**, 260-263.
*13.* McGuffin,L.J., Bryson,K., Jones,D.T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics* **16**, 404-405.
*14.* Jones,D.T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195-202.
*15.* SYBYL 6.9., Tripos Inc., 1699 South Hanley Rd., St. Louis, Missouri, 63144, USA.

## **PROFESY** - 70 models for 14 3D targets

### Protein structure prediction method based on fragment assembly and conformational space annealing

Julian Lee[1], Seung-Yeon Kim[2] and Jooyoung Lee[2*]
*[1]- Dept. of Bioinformatics and Life Science, Soongsil University,*
*[2] - School of Computational Sciences, Korea Institute for Advanced Study*
jlee@kias.re.kr

We have developed an improved version of PROFESY[1], a novel method for ab-initio prediction of protein tertiary structures based on fragment assembly and global optimization.

In contrast to the primitive version presented in CASP5, where the hydrogen bond was defined only in terms of inter-atom distances, its angle dependence is now incorporated. This new feature allows us to obtain low-energy conformations with a reasonable amount of beta strands, in contrast to the earlier version where the fraction of alpha helices was excessively large on average. In order to enhance the performance of the prediction method, we have optimized the linear parameters of an energy function, so that native-like conformations become energetically more favorable than non-native ones for proteins with known structures. The feasibility of the parameter optimization procedure is tested by applying it to the training set consisting of two proteins of the structural class $\alpha + \beta$: 1FSD and 1PQS. We use the resulting parameter set for jackknife tests, using several proteins from various structural classes. The results are quite promising. In particular, for protein 2GB1, the prediction results improve dramatically with the optimized the parameter set compared to the original parameters, despite the fact that it is *not included in the training set*. This suggests that parameters trained for a relatively small number of proteins are transferable to other proteins to some extent.

We have applied the PROFESY with the optimized parameters for the blind prediction of CASP6. The results will be discussed.

1. Lee,J., Kim,S.-Y., Joo,K., Kim,I., Lee,J. (2004). Prediction of Protein Structure Prediction using PROFESY, a novel method based on fragment assembly and conformational space annealing. *Proteins* **56**, 704-714.

## **ProteinShop** - 75 models for 15 3D targets

### Protein structure prediction using physics-based global optimization with knowledge-guided fragment packing

Jinhui Ding[1], Elizabeth Eskow[2], James Lu[1], Wei Liu[1,3], Lianjun Jiang[2], Richard Byrd[2], Robert Schnabel[2], and Silvia Crivelli[1,4]
*[1]California Institute for Quantitative Biomedical Research, Univ. of California, Berkeley, CA 94720, [2]Dept. of Computer Science, Univ. of Colorado, Boulder, CO 80309, [3]Dept. of Statistics, Univ. of California, Davis, CA 95616,*
*[4]Lawrence Berkeley Laboratory, Berkeley, CA 94720*
SNCrivelli@lbl.gov

We describe a protein structure prediction method that predicts the three-dimensional structure of new folds via minimizations of a physics-based energy function. The method is one of the few attempts to use an all-atom physics-based energy function throughout all stages of the optimization but it also uses filters to enhance the ability to discriminate among folds. It is based on the hypothesis that although the fold recognition servers can only provide limited and incomplete folding information for the targets in the new folds category,

that information may be valuable for guiding the global optimization process to find the native conformation.

Our method uses a novel fragment-assembly approach in which the structural fragments are constructed from the ideal geometric definitions of the local secondary structures using just sequence and secondary structure information. No structures of known proteins are used for the preparations of the structural fragments. The method is composed of two phases. Phase I creates an initial, extended configuration that has α-helices and β-strands according to the predictions. This configuration is split into fragments, each containing a single α-helix or β-strand and then the fragments are packed according to results obtained (if any) from the fold recognition meta-servers using the initial sequence of amino acids as a query. All the starting configurations are minimized locally to start the next phase. In phase II, both global and local optimization methods are applied to a number of the best minimizers generated in phase I. Phase II improves the initial configurations through global minimizations in subspaces of the dihedral angles of amino acids predicted to be coil.

**Method Description: Phase I**

In this phase, a variety of partially or fully folded initial configurations are constructed using secondary structure predictions. The predictions of secondary structure are primarily obtained from the *PSIPRED* server[1] but results from the *JUFO* server[2] are also considered. First, an unfolded configuration is created that has α-helices and β-strands according to those predictions. The extended configuration is created "from scratch" using *ProteinShop*[3], a manipulation tool that creates the three-dimensional coordinates of an extended protein structure containing α-helices and β-strands using sequence and predicted secondary structure information only. The extended configuration is divided into several structural fragments such that each structural fragment contains one rigid-body portion, which is either an α-helix or a β-strand. The cut point between the structural fragments lies in the region predicted to be "coil". The fragments are repacked using model templates. We use LiveBench[4], combined with 3D-Jury[5], to find the models (using the target sequence) and then we group the hits so that those hits that belong to the same SCOP family[6] are in the same group. A list of model templates is created by choosing those hits with the highest *3D-Jury* score in each group. In addition, several "welded" model templates, built by combining structural information from two hits from different groups, may also be included in the final list of model templates. Once the model templates are ready, the final set of initial configuration is constructed using one of the following approaches:

*Constructing Partially Folded Structures Using Templates:* One initial structure is built for each model template by packing the structural fragments according to the model template. For each structural fragment, one transformation matrix is calculated by aligning the rigid-body portion in the fragment to the corresponding portion on the template. The correspondence between the rigid-body portions on the target and those on the template is determined by the alignment generated by the meta-server. The transformation matrix is applied to the structural fragment for which the matrix was calculated. Often the model templates provide only partial information because of alignment gaps. If an alignment gap is in the middle of the sequence, the corresponding fragment is manipulated depending on the spatial limitations from the neighboring fragments. If the alignment gap is at the C- or N-terminals, the rigid-body fragments are connected extended. The coil regions between the two rigid-body portions are predicted by the loop prediction program developed by Xiang, *et al.*[7] whenever possible. When the loop regions cannot be predicted by the loop prediction program, we apply adjustments to each residue in the loop.

*Constructing Folded Structures Using Templates:* When there is no information available to model either the C- or N-terminals, we manually align the β-strands in the extended part with the β-strands in the folded part to form new β-sheets or to join existing β-sheets in the folded core. Depending on the number of the β-strands in the unfolded part, a variety of refined models are built from one partially folded model to create different β-sheet topologies. The fragment assembly and manipulation are performed by using *O*[8] and *ProteinShop*.

*Constructing Structures using BuildBeta: BuildBeta* is a *ProteinShop* feature that uses probability results on both protein fold topology[9] and sequence matching specificity[10] to automatically produce a high probability collection of possible initial sheet conformations. The current implementation of *BuildBeta* is limited to ten strands or less, which is the limit of Ruczinski's data fitting[9], and makes no attempt to make two or more sheets; all β-strands are placed in one β-sheet. Furthermore, it produces no result when the coil region between two β-strands is short (less than four residues). Usually, manipulations are necessary to avoid the severe steric overlap that may result after *BuildBeta*. The structures generated for CASP6 with *BuildBeta* accounted for about 8-20% of the initial structures created for Phase II.

**Phase II**

The second phase improves the initial structures by performing small-dimensional global minimizations in various subspaces of the parameter space followed by full-dimensional local minimizations. The method selects a number

of low-energy configurations from the list of initial structures and then selects small subsets for improvement by global minimizations. A subset of variables consists of a number of consecutive dihedral angles picked from the set of amino acids predicted to be coil by the secondary structure predictions. Once the subset is determined, a stochastic global optimization procedure is executed to find the best new positions for the chosen dihedral angles while holding the remaining dihedral angles fixed. A number of those configurations with the lowest energy values are selected for local minimizations in the full-dimensional space. The new full-dimensional local minimizers are then merged with those found previously, and the entire process repeats iteratively until the lowest energy configuration does not change substantially after a number of iteration steps.

Our method uses an all-atom AMBER force field with modified parameters[11]. The modified parameters are designed to improve the discriminatory ability of the energy function by enforcing the formation of hydrogen-bonds and β-sheets. Although our long-term goal in the context of *new folds* protein structure prediction is to find an effective energy function with some capability of distinguishing correct folds from misfolds, we believe a combination of molecular mechanics-based and protein database-derived potentials is the right direction to improve the ability to discriminate among folds in the short-term. This is particularly important for our method because it is computationally intensive. Thus, we began using some filters during Phase II to zero in on the most likely protein structures among the large number of potential candidates created by the minimization process. These filters evaluate certain attributes of the protein structures such as compactness, number of hydrogen-bonded pairs, and overall quality of the elements of secondary structure.


1. McGuffin,L.J., Bryson,K. & Jones,D.T. (2000) The PSIPRED protein structure prediction server. *Bioinformatics* **16**, 404-405.
2. Meiler,J., Mueller,M., Zeidler,A. & Schmaeschke,F. (2002) JUFO: Secondary Structure Prediction for Proteins. www.jens-meiler.de
3. Crivelli,S., Kreylos,O., Hamann,B., Max,N. & Bethel,W. (2004) ProteinShop: A tool for interactive protein manipulation and steering. *Journal of Computer-aided Molecular Design* **18**, 271-285.
4. Rychlewski,L., Fischer,D. & Elofsson,A. (2003) LiveBench-6: large-scale automated evaluation of protein structure prediction servers. *Proteins* **53** Suppl 6, 542-547.
5. Ginalski,K., Elofsson,A., Fischer,D., & Rychlewski,L. (2003) 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics* **19(8)**,1015-1018.
6. Murzin,A.G., Brenner,S.E., Hubbard,T. & Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536-540
7. Xiang,Z., Soto,C. & Honig,B. (2002) Evaluating Conformational Free Energies: The Colony Energy and its Application to the Problem of Loop Prediction. *Proc. Natl. Acad. Sci. USA* **99**, 7432-7437.
8. Jones,T.A., Bergdoll,M. & Kjeldgaard,M. (1990) O: A macromolecular modeling environment. In: Crystallographic and Modeling Methods in Mol. Design. Eds.: C. Bugg & S. Ealick. Springer-Verlag Press 189-195.
9. Ruczinski,I., Kooperberg,C., Bonneau,R. & Baker,D. (2002) Distributions of beta sheets in proteins with application to structure prediction. *Proteins* **48**, 85-97.
10. Zhu,H. & Braun,W. (1999) Sequence specificity, statistical potentials, and three-dimensional structure prediction with self-correcting distance geometry calculations of β-sheet formation in proteins. *Protein Sci.* **8**, 326-342.
11. Simmerling, C., Strockbine,B., & Roitberg,A.E. (2002) All-atom structure prediction and folding simulations of a stable protein. *J. Am. Chem. Soc.* **124**, 11258-11259.


## **Protfinder** (serv) - 250 models for 53 3D targets

### Protfinder, a physically motivated protein structure prediction algorithm


U. Bastolla[1]

[1] – *Centro de Astrobiologia (INTA_CSIC), Madrid, Spain*

bastollau@inta.es


The Protfinder algorithm predicts protein structures by aligning the query sequence to candidate structures in the PDB. Alignments are evaluated through a minimal model of protein folding, which reproduces approximately some key features of protein thermodynamics and is very convenient for rapid computation.

Information on sequence homology is not used in the scoring function. Nevertheless, when homologous proteins are present in the structure database, they are in almost all cases predicted as the best scoring structure.

Protein structures are represented as contact maps and their effective intramolecular interactions are modeled as a sum of contact interactions. The contact energy function used was derived in Ref.[1] through an optimization procedure, and assigns lowest energy to the experimentally known native structure for almost every sequence of monomeric protein whose structure has been determined by X-ray crystallography, except small fragments and chains with large cofactors. Moreover, it generates well-correlated energy landscapes, in the sense that structures very dissimilar from the native one have energies much higher than the native energy. This property is crucial for protein structure prediction. The effective free energy function is also able to estimate the folding free energies of a set of small proteins folding with two-state thermodynamics, with reasonable agreement with experimental data[2].

The scoring function consists of three elements: the effective energy function described above, a chain entropy term estimated in Ref.[2] and a term penalizing gaps in the alignment. Gaps in secondary structure elements are strictly forbidden. Gaps in the structure are allowed only if the two residues that are shortcut are close in space and the angles characterizing their pseudo-peptidic bond lie within a predefined range. Gaps in the sequence are allowed only on the surface of the protein, which is identified by the fact that the number of contacts per residue is smaller than a threshold. Allowed gaps receive an energetic penalty *G0* plus a penalty *G1* for each residue in the gap.

To speed up the computation, each structure in the NRDB90 non-redundant subset of the PDB was preprocessed to produce its contact map and the list of allowed shortcuts in the structure. Secondary structure was obtained from the DSSP file[3] when available, otherwise from the PDB file. The few structures for which no secondary structure assignment could be obtained were discarded. Preprocessing, together with the fact that the code uses mostly integer arithmetic, speed up considerably the computation.

To search for the optimal alignment, we use a stochastic version of the deterministic Build-up algorithm developed by Park and Levitt for searching low energy configurations of discrete protein models[4]. The algorithm is very efficient at finding high-scoring alignments, although it is not guaranteed to find the best optimum.

The algorithm starts by generating all possible gapless alignments of length *l* between the query sequence and the test structure and stores the M alignments with maximum score. At each subsequent step, an attempt is made to add a new residue to each of the M alignments. There are three possibilities: either the residue is aligned to the next structural position, or it is aligned introducing a gap in the structure (if allowed), or the residue is not aligned, initiating a gap in the sequence. All possible continuations are generated, and the M best scoring alignments are selected and used as seeds for the next step. The algorithm is iterated until no other residue can be added.

Some tricks are used to improve the efficiency of the algorithm: 1) The algorithm is first applied using a small value M=50 to scan rapidly the whole database. The 200 proteins with the best alignments are then stored in memory and used for a second more accurate search with M=800. 2) Instead of using the deterministic algorithm described above, we select the M alignments at each step based on the sum of their score plus a random number. The relative importance of the randomness is large in the first steps, allowing the algorithm to visit a larger fraction of the alignment space instead of constructing very similar alignments. The randomness decreases as the alignments get longer, so that the choice of the complete alignment is made on the basis of the deterministic score alone. 3) Since the construction of the starting fragment is the most delicate step, the algorithm is applied using two or three different values of the length of the initial fragment.

Each candidate structure receives the score of its best alignment. The best scoring structure is used as prediction. The goodness of the prediction is estimated through the normalized energy gap, a parameter measuring the difference between the best score and the score of an alternative structure in units of the best score, divided by the structural distance between the best scoring structure and the alternative structure. If the minimal value of the normalized energy gap over all alternative structures is large (larger than 0.2), the prediction is reliable, if it is small alignments with very different structure have scores quite similar to the best one and the reliability is very low.


1. Bastolla, U. et al. (2000) A statistical mechanical method to optimize energy functions for protein folding. *Proc. Natl. Acad. Sci. USA* **97**, 3977-3981
2. Bastolla, U. Testing the thermodynamics of a minimal model of protein folding, in preparation
3. Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22** (12), 2577-2637
4. Park, B.H. and Levitt, M. (1995) The complexity and accuracy of discrete state models of protein structure. *J. Mol. Biol.* 249, 493-507

## **PROTINFO** (serv) - 232 models for 64 3D targets

### Refining comparative models using a graph-theoretic approach

T. Liu and R. Samudrala
*University of Washington*
{tianyun,ram}@compbio.washington.edu

We evaluated the ability and effectiveness of a novel graph-theoretic approach to find the optimal interactions in a protein structure, given a variety of side-chain and main-chain conformational choices for each position. Sampling of side-chain and main-chain conformations was accomplished by exhaustively enumerating all possible choices from a population of initial models. The best combinations of these possibilities were selected through an all-atom scoring function[1] aided by the graph-theoretic approach[2].

For each CASP6 target, several models were generated using 3D-Jury server (http://BioInfo.PL/Meta)[3] combined with our comparative modeling server, PROTINFO-CM (http://protinfo.compbio.washington.edu)[4]. Additional models were obtained from the CAFASP4 server after scrutinizing the alignments to gain extra variability in sequence alignments and templates. Models were inspected for missing or incorrect parts, typically for loops. If reasonable alternative loops could be built using our in-house software, they were added to the pool as well. Side-chain possibilities were also constructed using the program SCWRL[5]. Care was taken to assure that models were superimposed based on their secondary structure so that the average α-carbon root mean square deviation (cRMSD) between each model was less than 5 Å.

After a set of models was superimposed, the next step required the determination of the crossover points where mixing between different parent structures could occur. Crossover points were defined by the ranges of main-chain where the α-carbon was less than 1.0 Å from each other, and were not permitted inside secondary structure elements.

We then used a graph-theoretic clique-finding approach to assemble the sampled side-chain and main-chain conformations. A complete description of the method is given elsewhere[2]. The idea of this approach is to obtain optimized mosaic models by shuffling them in a rational way. Thus the key point is the choice of an appropriate scoring function. We used an all-atom conditional probability discriminatory function (RAPDF)[1] to evaluate the cliques, with the highest scoring ones representing the optimal combinations of the different side-chain and main-chain possibilities.

In the final step, all models from the above approach were refined with ENCAD[6]. The effectiveness of this methodology to improve the model accuracy remains to be investigated.

1. Samudrala,R., Moult,J. (1998) An all-atom distance dependent conditional probability discriminatory function for protein structure prediction. *J Mol Biol* ***275***, 893-914.
2. Samudrala,R., Moult,J. (1998) A graph-theoretic algorithm for comparative modelling of protein structure. *J Mol Biol* ***279***, 287-302.
3. Ginalski,K., Elofsson,A., Fischer,D., Rychlewski,L. (2003) 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics*. ***19***: 1015-1015.
4. Hung,L-H., Samudrala,R. (2003) PROTINFO: Secondary and tertiary protein structure prediction. *Nucleic Acids Research* ***31***, 3296-3299.
5. Bower M.J., Cohen F.E., Dunbrack R.L. (1997) Prediction of side-chain orientations from a backbone-dependent rotamer library: A new homology modelling tool. *J Mol Biol* ***267***, 1268-1282.
6. Levitt,M., Hirshberg,M., Sharon,R., Daggett,V. (1995) Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. *Comp Phys Comm* ***91***, 215-231.

## **PROTINFO-AB** (serv) - 160 models for 32 3D targets

### Generating, selecting and refining protein structures *de novo*

L-H. Hung, S.C. Ngan, and R. Samudrala
*University of Washington*
{lhhung,ngan,ram}@compbio.washington.edu

We have implemented a new tri-partite protocol for the automated prediction of protein structure from sequence alone. Structures are generated using a simulated annealing search phase that minimizes a target scoring function. Moves are derived from a synthetic function that produces $\phi/\sigma$ angular distributions similar to the empirically observed ones. In contrast to fragment based methods, this is accomplished without copying any angles or coordinates. After the search phase, a local minimization protocol further reduces the target score. In cases where there are strands or constraints, a pre-condensation phase

allows strands to pair and constraints to be satisfied. A series of composite functions based on different combinations of 14 individual scoring functions is used to choose a set of best conformers. A novel iterative density protocol is then used to choose the best structures from this set. Finally, the best conformers are used to guide the generation of new conformers, thus iteratively refining the predicted structure. As of this abstract submission, T0236 model 5 is an example of the protocol where a structure better than most of the fold recognition models (cRMSD of 1.97 A for residues 1-50, 5.42 A for all 84 residues) is produced.

Generation of structures is accomplished through a search phase where a composite energy function is minimized by Monte-Carlo simulated annealing. In contrast to methods that replace fragments from known structures, the present protocol uses a function that generates $\phi/\sigma$ angles that reflects the distribution observed in the PDB, and does not copy any angles or coordinates. All residues in a given protein sequence are first classified by the encompassing triplet sequence and the triplet secondary structure. A histogram is then constructed from the $\phi/\sigma$ angles of matching triplets of the same secondary structure in the PDB. (A bin size of 10 degrees by 10 degrees is used and only the angles in the central residue of the triplet are plotted). The mean $\phi/\sigma$ angle in each bin and the standard deviation are recorded. To choose a $\phi/\sigma$ pair during the simulation, a bin is first chosen using the frequencies observed in the histogram. The angles are then chosen using a normal distribution that fits the mean and standard deviation of the observed distribution within the bin.

In addition to the main search phase we have also added a minimization phase using Brent's method and small random moves which typically result in a further 10% reduction in the target score. A pre-condensation phase, implemented late in CASP, encourages pairing of strand residues and satisfaction of other constraints resulting in 10-100 fold increase in the number of paired strands formed.

The search target function is a compromise between the speed of evaluation and the best correlation to the distance from the native structure. We keep the 10 best conformers per seed for analyses using 14 energy-like scoring functions encompassing physical energy functions (*vdw*, *electrostatic*, *solv*), general empirical functions (*Shell*, *MJ*, *hcf*, *Sol*, and *Rad* ) and PDB-based empirical functions (*RAPDF*, *Coord*, *Conseq* and *Curv*). Due to the diversity of both the functions and the proteins that are being evaluated, it is difficult to derive a single weighting scheme that produces an optimal composite function. Instead, the best linear combinations of these functions were determined by logistic regression on large sets of decoys. 19 groups of these linear combinations were used to filter the initial set of conformers. Typically, 100,000 – 200,000 conformers are reduced to about 1000-2000 at this stage

Energy-like scoring functions alone are still very inconsistent at picking out the best structures. Fortunately, one of the most powerful scoring functions is the completely statistical density function, which is the (negative) total RMSD to the other conformations in the set and is a measure of the distance of a conformer to the center of the distribution. Unfortunately, the largest contributions to the density scores come from the outliers that can skew the correlation of density to the distance to the true center of the distribution, reducing the effectiveness of the function. Thus, we have implemented a new iterative density function that measures the density, removes the worst outlier (the conformer with lowest density) and then repeats the process until there are no more outliers in the set. The center of this trimmed set is then selected (and the centers of the largest k-means clusters for the final selection of 5 for CASP) and is taken as the best.

Finally, if there is a good cluster of conformers it is possible to generate a better set of conformers near the conformational center. This is done by incorporating the RMSD to the best conformers into the target function and/or using internal distance constraints derived from these conformers and repeating the generation stage. Selection of the best conformers proceeds as before and the spread of the final set of 5 conformers is reduced to 2-4 A cRMSD.

**PSWatch** - 3 models for 3 3D targets

## Protein Structure Watch: making “predictions” easy

A.G. Murzin
*MRC Centre for Protein Engineering*
agm@mrc-lmb.cam.ac.uk

Protein Structure Watch is an informal protocol of gathering intelligence on new protein structures asap, usually well in advance of official publication. Developed as a SCOP pre-classification tool, this approach exploits new trends in structural biology observed during classification of recent structures. Two factors are of particular significance for CASP present and future. One is the changing attitude of structural biologists, who now are keener than ever to advertise and to publish their new structures sooner, the other is the increasing duplication of their efforts, resulting in more than one structure for almost every new protein family determined independently at about the same time.

This approach has been adapted for CASP6 to find publicly available information on the structures of targets and their probable close and distant homologues, focusing mainly on the targets without a close homologue in PDB at the beginning of the prediction season. The main goal of this exercise was to estimate a combined damage to the pool of such targets from the leaks of structural information and duplicated structure determinations. An additional goal was to demonstrate the potential effect of privileged access to unpublished structures. Many, but not all, findings have been documented in my comments to CASP6 targets on the FORCASP site and are summarized in my CASP6 “Methods” paper there.

**Pushchino** - 194 models for 62 3D targets

## Threading the use of multiple homology and secondary structure prediction information

M.Yu. Lobanov[1], D.N. Ivankov[1], S.O. Garbuzynskiy[1], N.S. Bogatyreva[1], O.V. Galzitskaya[1], I.I. Litvinov[2], M.A. Roytberg[2], A.V. Finkelstein[1]
[1]*Institute of Protein Research RAS*
[2]*Institute of Mathematical Problems in Biology RAS*
afinkel@vega.protres.ru

For creating bunches of reliable homologous sequences we used PSI-BLAST[1]. Secondary structure of targets was predicted by PsiPred[2]. Secondary structure of 3D templates was calculated by DSSP[3].

To divide the target by domains we used alignments of HMMer SUPERFAMILY[4], alignments of PSI-BLAST and our program[5] (see abstract of group “Oka”).

Threading was done by our program SCF_THREADER[6] with the scoring function that takes into account the following factors: similarity of sequences (by GON250), similarity of secondary structures (see table below), 3D-structure dependent gap penalties, 3D constrains of gaps in sequences threaded onto a template.

| | | **DSSP** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **H** | **G** | **I** | **E** | **B** | **S** | **T** | **-** |
| **PsiPred** | **H** | -0.85 | -0.04 | -0.04 | 3.41 | 2.19 | 2.18 | 1.09 | 2.4 |
| | **E** | 3.44 | 1.72 | 1.72 | -1.45 | 0.02 | 1.64 | 2.84 | 0.86 |
| | **C** | 2.27 | -0.31 | -0.31 | 0.98 | -0.65 | -0.83 | -0.76 | -0.79 |

For evaluation of the results we used all threading parameters stated above and energies of long range contacts. Besides, in some cases we also used results of SUPERFAMILY and PSI-BLAST servers to select the best templates. Finally, a visual inspection of the best results presented by the program SCF_THREADER was done to reject incompact structures, as well as structures with a single β-strand or other structural defects.

The current version of program SCF_THREADER differs from the previous one in CASP5 by (1) more detailed parameters for secondary structure

comparison, (2) more detailed 3D gap penalties, (3) the use of long range interaction energies for evaluation of the final alignments.


1. Altshul,S.F. et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25** (17), 3389-3402.
2. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292** (2), 195-202.
3. Cabsch,W., Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers.* **22** (12), 2577-637.
4. Gough,J., Karplus,K., Hughey,R., Chotia,C. (2001) Assignment of homology to genome sequences using a library of hidden markov models that represent all proteins of known structure *J. Mol. Biol.* **313** (4), 903-919.
5. Galzitskaya,O.V. & Melnik,B.S. (2003). Prediction of protein domain boundaries from sequence alone. *Protein Sci.* **12**, 696-701.
6. Rykunov,D.S. (2000) Search for the most stable folds of protein chains: III improvement in fold recognition by averaging over homologous sequences and 3D structures. *Proteins* **40** (3), 494-501.


## **RAGHAVA-GPS** - 124 models for 60 3D / 63 FN targets

### Prediction of Genome Ontology (GO) class of a protein from dipeptide composition Gaussian method


G. P. S. Raghava
*Institute of Microbial Technology*
*Sector-39A, Chandigarh, INDIA*
raghava@imtech.res.in


Functional annotation of proteins is one of the major challenges in era of genomics, as number of proteins whose sequence is known is growing with exponential rate due to advancement in DNA sequence techniques. Even the genomes of important pathogens like M. Tuberculosis, is partially annotated, most o the proteins are assigned theoretical proteins. Though attempts have been made in past to predict function but progress performance is not very high. In this study we have made an attempt to predict class of proteins as per Genome ontology (GO) classification. Genome ontology is one of the major source of information from where one can obtained the information of class of protein. In GO database the annotation of proteins are at three level i) Biological functions; ii) Biological Process and iii) cell. However, a large number of method already developed in past to predict the class of proteins are limited to predict few classes of proteins. In this study we create the dataset of proteins for each class of GO. It was observed that most of the GO class have very limited number of proteins thus it is difficult to develop rediction method for these classes. In order to avoid this problem we only keep families which have 50 or more proteins. These proteins were obtained from UNIPROT database where function of these proteins is manually annotated as per GO classification. We trained our method using Gaussian technique available in LNKNET software. The dipeptide compostion was used as input pattern of proteins. In order to predict the functional class of a query sequence (CASP6 targets), first dipeptide composition of query sequence is calculated then we predict the class of protein using rules derived for each class using Gaussian routine of LNKnet software.

Tertiary structure of proteins using a de novo method design for prediction of small bioactive peptides.

We developed a method for predicting tertiary structure of bioactive peptides. The tertiary structure prediction of such peptides can aid in understanding of biological function and protein structure prediction. Our strategy for prediction of tertiary structure of small peptides is based on the observation that β-turn is an important and consistent feature of small peptides in addition to regular structures. It has been found that 75.3% of total peptides analyzed in present study have at least one β-turn. For this reason, it should be possible, given their sequences, to make accurate predictions about their structure using both the regular and irregular secondary structure information, mainly of β-turns. Thus regular and irregular secondary structures, particularly β-turns information can play a vital role in prediction of tertiary structure of small bioactive peptides. A representative data set comprising of three-dimensional structures of 77 biologically active peptides have been obtained from PDB[1] and other databases such as PSST (http://pranag.physics.iisc.ernet.in/psst) and PRF (http://www.genome.ad.jp/). The data set has been restricted to those biologically active peptides that consist of only natural amino acids and are linear with length varying between 9-20 residues. We have analyzed these 77 biologically active peptides. Out of 77 peptides, 58 peptides have been found to contain at least one β-turn. At residue level, about 34.9% of total peptide residues fall in β-turns, higher than the number of helical (32.4%) and β-sheet residues (6.9%). Based on these observations, four different models have been generated using predicted secondary structure information. The first model I, has been build up by assigning all the peptide residues the extended conformation ($\phi = \Psi = 180^{o}$). Second model II, has been build by using the

information of regular secondary structures (helices, β-strands and coil) predicted from PSIPRED[2]. In third model III, secondary structure information including β-turn types predicted from BetaTurns has been used[3,4]. The fourth model IV has main-chain ϕ, Ψ of model III and side chain angles assigned using standard Dunbrack backbone dependent rotamer library. The models have been refined further by energy minimization with dynamics simulations using AMBER version6. It has been noted that the backbone averaged rmsd values before and after energy minimization are 10.8Å, 7.8Å, 5.5Å & 5.5Å and 6.4Å, 5.0Å, 4.4Å and 4.3Å for models I, II, III and IV respectively. The results indicate that secondary structures, particularly β-turns can provide valuable information for tertiary structure prediction. Based on above study, we have developed a web server PEPstr which allows the tertiary structure prediction of small bioactive peptides using the following steps i) prediction of regular secondary structure and β-turns using BetaTurns; ii) generation of conformation by assigning dihedral angles corresponding to secondary structure information; iii) placement of side chain angles using Dunbrack backbone dependent rotamer library; and iv) energy minimization using AMBER. The server Pepstr is accessible from http://www.imtech.res.in/raghava/pepstr/. In CASP6 we used above method for predicting structure of protein.


1. Bernstein,F.C., Koetzle,T.F., Williams,G., Mayer,E.F., Bryce,M.D., Rodgers,J.R., Kennard,O., Simanouchi,T. and Tasumi,M. (1977) The Protein Data Bank: a computer based archival file for macromolecular structures. *J. Mol. Biol*. **112**, 535-542.
2. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol*. **292**, 195-202.
3. Kaur,H. and Raghava,G.P.S. (2003) Prediction of β-turns in proteins from multiple alignment using neural network. *Protein Sci*. **12**, 627-634.
4. Kaur,H. and Raghava,G.P.S. (2004) A neural network method for prediction of β-turn types in proteins using evolutionary information. *Bioinformatics* (in press).


## **RAGHAVA-GPS-mango** (serv) - 171 models for 57 FN targets

### MANGO: prediction of Genome Ontology (GO) class of a protein from its amino acid and dipeptide composition using nearest neighbor approach


G. P. S. Raghava
*Institute of Microbial Technology*
*Sector-39A, Chandigarh, INDIA*
raghava@imtech.res.in


One of the major challenges in era of genomics is to predict the function of proteins. As number of proteins whose sequence is known is growing with exponential rate due to advancement in DNA sequence techniques. This has pose a major challenge to the boinformatician to develop strategy to predict the function of protein. Fortunately, function of a large number proteins have been deduced using experimental techniques, one may obtained the information about manually annotated proteins from SWISSPROT database. Recently initiatives were taken to provide the uniform definition of class of protein. Genome ontology is one of the major source of information from where one can obtained the information of class of protein. In GO database the annotation of proteins are at three level i) Biological functions; ii) Biological Process and iii) cell. However, a large number of method already developed in past to predict the class of proteins are limited to predict few classes of proteins. In this study we create the dataset of proteins for each class of GO. These proteins were obtained from UNIPROT database where function of these proteins is manually annotated as per GO classification. For each class of GO we create the average composition of proteins belongs to that class. Lets a given GO class have 200 proteins than we compute overall composition of each of 20 the natural residues. This residue composition represents the class. In order to predict the functional class of a query sequence (CASP6 targets), first composition of query sequence is calculated then we compute the Euclidian distance between composition of query sequence and each class of GO. The class having minimum Euclidian distance were assigned as class of query proteins.

It has been shown in past that dipeptide composition have more information than simple composition because order of neighbor is also considered. Thus we implement our approach using dipeptide composition, where dipeptide composition of proteins were used to calculate Euclidian distance between query protein and GO class of proteins instead of residue composition. We also

compute the overall difference (residue composition and dipeptide composition) in query and GO class of proteins. In summary we used composition, dipeptide composition and comination of both for predictiog GO class of target proteins.

## **RAGHAVA-GPS-rpfold** (serv) - 165 models for 48 3D targets

### A server for predicting fold of protein from its sequence using secondary structure and PSIBLAST profile

G. P. S. Raghava
*Institute of Microbial Technology*
*Sector-39A, Chandigarh, INDIA*
raghava@imtech.res.in

A web server RPFOLD has been developed for searching known fold in protein. This server uses the freely available tools for prediction like PSIBLAST, PSIPRED, SSEARCH etc. First sequence similarity search is performed using SSEARCH (FASTA Package) where query sequence is searched against all the protein sequence in protein data bank (PDB. Thus we obtain set **A** of proteins whose structures are known and have similarity with query sequence. All the proteins in set **A** have similarity score obtained from SSEARCH. In next step we performed similarity search using sequence profile. First query protein sequence was searched against non-redundant database using 3 iterations of PSIBLAST and profile was generated. Thus we obtained sequence profile corresponding to query proteins. The sequence profile of query protein was used as input to perform sequence similarity search against proteins in PDB. This way protein that have remote similarity were also obtained. We create set **B** of proteins who have similarity with query sequence and were obtained from sequence profile search using PSIBLAST. In third step we performed structure similarity search instead of sequence similarity search. In this case first we obtained secondary structure of all protein in PDB where secondary structures were assigned using DSSP then we predict the secondary structure of query protein using PSIPRED. Predicted secondary structure was searched using FASTA against database of secondary structure of proteins in PDB. We create set **C** of proteins who have secondary structure similarity with secondary structure of query proteins.

This way we obtain three sets (**A, B & C**) of proteins of known folds who have similarity with query sequence obtained using different similarity search criteria. All the proteins in three sets have similarity scores. Finally we combine three sets and ranked based on score and weightage. Clustal-W was used to align query sequence with predicted secondary structure information and target protein in PDB with assigned secondary structure information to get final alignment and re-ranking of hits.

## **RANKPROP** - 64 models for 64 3D targets

### Fold recognition by protein ranking on the protein similarity network

Martial Hue[1], William Stafford Noble[2] and Jean-Philippe Vert[1]
[1]- *Centre de Geostatistique, Ecole des Mines de Paris,*
[2]- *Departement of Genome Science, Universiyt of Washington*
hue@cg.ensmp.fr, noble@gs.washington.edu, Jean-Philippe.Vert@mines.org

**1. Introduction**

Our participation to CASP focuses on the Comparative Modeling and Fold Recognition tasks. For each target we submitted a list of putative homologs with known structure aligned to the target. The originality of our approach is the method used to predict putative homologs from sequence information only.

We implemented the method of Noble et al.[1] that, for each target, performs a ranking, by decreasing similarity, of all sequences in a large database using the global structure of the protein similarity network. This method was shown to outperform PSI-BLAST for the recognition of homologies at the SCOP super-family level, and our main motivation was to test its relevance for the problem of structure prediction when homology recognition at the super-family level is useful, i.e., mainly for targets classified as “Fold Recognition” targets.

**2. Method**

Given a target sequence, the goal is to prepare a prediction file in the AL format, i.e., to propose a list of sequences in PDB aligned to part of the target sequences (with no overlap between the proposed alignments). Our approach consists in 1) computing a ranking of a large set of protein sequences with respect to the target using the global protein similarity network, 2) reading this ranking from top to bottom, and 3) for each sequence with known structure (i.e., in PDB), aligning the target sequence using a hidden Markov model containing the hit sequence (with the HMMER software).

Description of the Protein Similarity Network. The Protein Similarity Network used was obtained from the Biozon database (www.biozon.org) provided by

Golan Yona. This network can be logically viewed as a weighted graph, with vertices corresponding to protein sequences and weighted edges corresponding to similarities, more precisely E-values computed by the program BLAST. The list of sequences on which the computation is performed contains a large number of amino acid sequences, 933075 at the beginning of the CASP competition, 933116 at the end, since some of the targets did not belong to the graph. They are numbered from 1 to 933116. The whole graph is distributed in 94 files, for a total of roughy 30 gigabytes. Each query sequence of our ranking algorithm requires the query to be a node in the graph, so each time a CASP target was not already in the graph, we first added it.

Algorithm. We implemented the algorithm described in Noble et al. (2004) to rank the sequences in the network with respect to a given query. Let n be the number of vertices of the graph, i the index of the target sequence, W the similarity matrix, of size n*n, defined by

$$W_{ij} = \exp(-Evalue_{ij}/\sigma)$$ .

it is a number between 0 and 1. If the E-value is small, the similarity is closer to 1. $\sigma$ has the role of a threshold on the weights of the edges : E-values larger than $\sigma$ are negligeable. S is the normalized similarity matrix, so that the sum of each line is 1, i.e.

$$S = D^{-1}W$$

where D is a diagonal matrix, with $D_{ii} = \Sigma_j W_{ij}$. Given a target i, let us define Y a n-vector equal to 1 on the i-th component, and 0 elsewhere. We perform 10 iterations of:

$$F(t+1) = \alpha S F(t) + (1-\alpha) Y$$

with $F(0) = Y$. $\alpha$ is a parameter between 0 and 1. Therefore for each t, F(t) sums to 1. If $\alpha = 1$, this is a diffusion. If $\alpha = 0$, at each iteration, F(t) is constant, equal to Y. In the experiments we took
The scores yielded by this “diffusion” are sorted; hence we have a ranking of the vertices of the graph.

Submission file. To make the submision file in AL format, the next steps are :
-- Extract the ranked sequences from SCOP/PDB (usually SCOP domains)
-- Define a vector U of bits, of the length of the target sequence, corresponding to the aligned letters so far.
-- Take the best ranked of these, and align the target onto it using HMMER. HMMER has an implementation of profile hidden Markov models for all SCOP superfamilies that allow robust alignments of a sequence to a superfamily.
-- Take the next best ranked SCOP entry (that belongs to a different superfamily from the first one) and align it also, and check that the positions aligned are different from the positions aligned with the first hit.
Only alignments of 8 consecutive letters in the target are accepted, with possibly a gap of 1 letter.
-- Iterate if necessary to find all domains, until the 2000-th sequence of the ranking.
-- Output the concatenation of these alignments and submit the file.

Parameters. The algorithm depends on two parameters $\alpha \in (0,1)$ and $\sigma$. The values chosen were $\alpha = 0.95$ and $\sigma \in \{1e\text{-}4, 0.01, 0.1\}$.

Conclusion. Our main motivation was to test the relevance of the new ranking algorithm for the detection of remote homologies. We will therefore check in details the results obtained for the “fold recognition” targets, which typically require efficient recognition at the super-family level. Our method is fully automatic and no human processing was performed; we did not participate as a server simply because the server was not ready at the beginning of the competition.


1. Weston,J., Elisseeff,A., Zhou, D., Leslie,C.S. & Noble,W.S. (2004). Protein Ranking : From local to global structure in the protein similarity network. *Proceedings of the National Academy of Science* **101 (17)**, 6559-6563.
2. Zhou,D., Weston,J., Gretton,A., Bousquet,O., Schoelkopf,B. (2003) Ranking on data manifodls. *Proceedings of NIPS.*


## **RAPTOR** (serv) - 292 models for 64 3D targets

### Regression-based approaches to fold recognition


Jinbo Xu and Ming Li
*School of Computer Science, University of Waterloo*
{j3xu, mli}@uwaterloo.ca


Protein structure prediction by protein threading technique has demonstrated a great success in recent CASPs (Critical Assessment of Structure Prediction) Protein threading makes a structure prediction by finding the optimal alignment between the target sequence and each of the available protein structures (also

called templates) in Protein Data Bank (PDB), and then choosing the best overall template as the basis on which the structure of the target sequence is built. The algorithm for finding the optimal sequence-template alignment has been researched extensively. However, how to choose the best template based on alignments is also critical to the success of protein threading. Fold recognition requires certain criteria to identify the best template for one target sequence. The sequence-template alignment score cannot be directly used to rank the templates due to the bias introduced by the residue composition and the number of alternative sequence-template alignments. So far, there are two strategies used by the structure prediction community for fold recognition: recognition based on Z-scores, and recognition by machine learning methods. Most of the current prediction programs use the traditional Z-score to recognize the best-fit templates, whereas several programs such as GenTHREADER and PROSPECT-I use a neural network to rank the templates. The neural network method treats the template selection problem as a classification problem. Z-score was proposed to cancel out the bias caused by sequence residue composition and by the number of alternative sequence-template alignments.

The Z-score method has the following two drawbacks: (i) It takes a lot of extra time to calculate Z-scores, especially the alignment number-corrected Z-scores. In order to calculate the alignment number-corrected Z-score for each threading pair, the target sequence has to be shuffled and threaded many times. In order to save time, many prediction programs like PROSPECT-I only calculate the composition-corrected Z-score. Even though this, the computational efficiency hinders the Z-score method from genome-scale structure prediction. (ii) Z-score is hard to interpret, especially when the scoring function is the weighted sum of various energy items such as mutation score, environmental fitness score, pairwise score, secondary structure score, gap penalty and score induced from NMR data. For example, when the sequence is shuffled, shall we shuffle the position specific profile information and the predicted secondary structure type at each sequence residue? If we choose to shuffle the secondary structure, then the shuffled secondary structure arrangement does not look like a protein's. Otherwise, if we choose to predict the secondary structure again, the whole process will take a very long time.

In our previous paper, we have very briefly introduced the SVM classification method for fold recognition. Although classification-based methods run much faster and have better sensitivity than the Z-score method, they still have some problems. The similarity between two proteins could be at fold level, superfamily level or family level. Classification-based methods can only treat the three different similarity levels as a single one. Multi-class SVM cannot be used here since the relationship among the three similarity levels is hierarchy. That is, if two proteins are in a family, then they are also in a superfamily and have the same fold. Classification-based methods cannot effectively differentiate one similarity level from another. The other problem is that even if SVM classification can predict two proteins to be similar in at least fold level, it is possible that the alignment accuracy between them is really bad. A template with only the same fold as the target sequence might be ranked higher than a template in the same family as the target, which is not what we expect.

We have developed two regression-based approaches (SVM regression and Gradient Boosting) to directly predict the alignment accuracy of a given sequence-template alignment. The predicted alignment accuracy has a high correlation coefficient with the real alignment accuracy. Then, we use the predicted sequence-template alignment accuracy to rank all the templates for a given sequence. Experimental results show that the predicted alignment accuracy has a much better sensitivity and specificity than composition-corrected Z-score method and a much better computational efficiency. Both regression-based methods are also better than other classification and the alignment number-corrected Z-score methods in terms of sensitivity. In addition, The alignment accuracy is also easier to interpret than the classification results.


1. Xu,J. (2004) Protein Fold Recognition by Predicted Alignment Accuracy. Submitted to IEEE Trans. On Computational Biology and Bioinformatics.
2. Jiao,F. and Xu,J. (2004) Protein Fold Recognition Using Gradient Boosting. Submitted to *Bioinformatics*.


## **Rohl** - 78 models for 51 3D targets

### Generating optimal 3D models from Robetta alignments


F.D. Khatib, J. Samayoa, D.L. Bernick, C. Lowe, C. Gorringe and C.A. Rohl

*University of California at Santa Cruz, USA*

rohl@ucsc.edu


Our structure predictions for CASP6 focused on the problem of using a given alignment to construct an optimal three-dimensional model of a query. In particular, we were interested in assessing methods that could be incorporated into the comparative modeling strategy utilized by the automatic structure prediction server, Robetta[1]. Consequently, we restricted our predictions to

those query sequences for which the Robetta server made predictions using a parent of known structure, and in general, we limited our consideration of possible parent structures and alignments to the parents and alignments reported by the Robetta server. Because we relied solely on the parent structures and alignments reported by the Robetta server, the quality of our models are expected in large part to be determined by the quality of those alignments. Consequently, we are interested primarily in assessing the quality of our predictions relative to those of the Robetta server. Given a particular parent and limited set of possible alignments, were we able to generate a higher quality model than the automated server? While our approach is similar to the homology-based modeling strategy employed by the Robetta server, significant differences include: 1. re-ranking of Robetta alignments using a consensus scoring method, 2. application of a filter for detecting knotted structures, 3. increased sampling of conformations for structurally variable regions corresponding to gaps in the alignment, 4. *de novo* construction of initial conformations for long internal gaps using only local structural information from the fixed template, 5. refinement of the models by optimization of an all-atom energy function, and 6. manual intervention for some targets.

The initial step in our protocol was to reevaluate the alignments utilized by the Robetta server to attempt to improve the relative rankings of these alignments. Each of alignments reported by Robetta for its top five predictions was assigned a score equal to the total number of occurrences of every aligned residue pair in any of five alignments or in the default alignment provided by the K*sync algorithm[2]. The score for each alignment was normalized by the number of aligned pairs and then used to rank the alignments. In general, only the top-ranked alignment was considered for further modeling. In cases where the top two or three alignments had comparable scores, each of the alignments were used for further modeling and the final selection was made on the basis of manual assessment of final model quality. Given an alignment, aligned residues were modeled by extracting coordinates of corresponding residues from the parent structure to generate a fixed template structure. Conformations for gapped regions were constructed using fragment-assembly protocols within the Rosetta structure prediction suite that have been adapted to model structurally variable regions in homologous proteins[3]. An initial library of conformations for gapped regions 17 residues or shorter were selected from a database of protein structures on the basis of sequence and secondary structure similarity and geometric fit to the template. Initial conformations for internal gapped regions 12 residues and longer were constructed *de novo* by fragment assembly in the presence of only sequentially adjacent template residues to ensure geometric fit. When sufficient initial conformations that met the steric constraints of the template could not be identified or constructed, template regions were manually trimmed back to enlarge the gapped region. Typically, such regions represented deletions relative to the parent structure. These initial conformations for internal gapped regions were screened for steric clashes with the template and knots, and then randomly assembled onto the template structure. Unaligned terminal segments were added to the models by fragment assembly and the resulting models were then optimized for agreement with the atom-based Rosetta energy function using both modified fragment replacement and small perturbations of torsion angles in both aligned and unaligned regions[4]. In some cases, as noted in the REMARKS section of the predictions, models were manually sculpted using the Protein Shop package[5] and then re-optimized within Rosetta. Other instances in which alternate strategies were undertaken for particular targets are also noted in the remarks section of the submitted models.


1. Kim,D.E., Chivian,D., Baker D. (2004) Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res*. **Suppl 2,** W536-31.
2. Chivian,D., Kim,D.E., Malmstrom,L. (2003) Automated prediction of CASP-5 structures using the Robetta server. *Proteins* **53** Suppl 6**,** 524-33.
3. Rohl,C.A., Strauss,C.E.M., Chivian,D., Baker,D. (2004) Modeling Structurally Variable Regions in Homologous Proteins using Rosetta. *Proteins* **55,** 656-677.
4. Rohl,C.A. (2004) Structure Estimation from Minimal Restraints Using Rosetta. *Methods in Enzymsology.* In press.
5. Kreylos,O., Max,N., Crivelli,S. (2002) ProtoShop: Interactive Design of Protein Structures, in: Moult,J., Fidelis,K., Zemla,A, Hubbard,T. eds. *Proceedings of CASP5 - Fifth Meeting on the Critical Assessment of Techniques for Protein Structure Prediction.* A213-A214.


## **Rokko** - 228 models for 64 3D targets

### De novo structure prediction by the SimFold energy function with the multi-canonical ensemble fragment assembly


Y. Fujitsuka[1*], G. Chikenji[2*], S.J. Park[3], W. Jin[2], N. Koga[1], T. Furuta[2], and S. Takada[1,2]

*[1] – Grad School, Sci & Tech Kobe Univ, [2] – Faculty of Sci, Kobe Univ, [3]- Interdisciplinary Grad School of Sci & Eng, Tokyo Inst Tech*

stakada@kobe-u.ac.jp


The team Rokko primarily focuses on *de novo* structure prediction for targets that possibly have “new folds”. For all targets, we first refer the results of PDB-

BLAST(4), many CASP servers, and 3D-jury(5) to filter out those that are likely to have good structural templates. For targets that might have "new folds", we prepare the fragment candidates of every 10 residues and perform multicanonical ensemble fragment assembly simulation(3) using our in-house developed energy function, SimFold(1,2). Models are chosen by clustering low-energy structures and by human inspection.

Here, we briefly describe 1) generation of fragment candidates, 2) energy function SimFold, 3) conformational sampling by the multi-canonical ensemble fragment assembly (FA), and 4) how we did in CASP6 including domain parsing, final model selection, and models for CM & FR targets.

1) Generation of fragment candidates: For the query sequence, the n-residue (n=10, 16, 21) PSSMs of PSI-BLAST are used to retrieve fragment candidates of every 10 residues from 2100 proteins that have known structures and share <20% sequence identity. For retrieval, two scores, the correlation coefficient and a Kurtosis-weighted dot product are used. Comparing the consensus of secondary structure predictions with the secondary structures of templates, we prepare 100 initial fragments for each 10 residues of the target, which is followed by reduction via clustering. As a result, the number of fragments for conserved sites is small, whereas that for the diverse site is large.

2) SimFold, the energy function(1,2): The protein is represented by a coarse-grained model, in which side chain atoms are replaced by a center of interactions. The interaction potential, which we call SimFold contains van der Waals interaction, secondary structure propensity, the hydrogen bond interaction, the hydrophobic interaction and the pair-wise interaction. The latter three depend on degree of burial of interacting atoms. No protein specific potential such as secondary structure prediction based potential is used in the energy function. Parameters in SimFold are optimized by Z-score optimization method.

3) Multicanonical ensemble fragment assembly: For conformational sampling, we use a variant of fragment assembly (FA) method called "reversible FA method" which we have recently developed (an earlier version in ref.3). Our FA is different from what has been developed by Baker's group. The most important difference between conventional FA and ours is that the conventional FA protocol does not fulfill the detailed balance condition, but our algorithm does. Thanks to this property, we could combine reversible FA with the multi-canonical ensemble Monte Carlo method which is known to be highly powerful conformational sampling method for protein systems. Indeed, this approach is used in all targets that are likely to have "new folds" helping conformational sampling very significantly especially for longer targets.

4) How we did in CASP6:

a) Domain parsing by Donuts (DOmaiN parsing UTility Software):Donuts is a tool that parses domain regions from an amino acid sequence. First, it searches any homologs in structural database with PDB-BLAST followed by 3D-Jury. For part of sequence that has no homologs, search is done with rpsblast against the Conserved Domain Database(6,7). Finally, the method of Galzitskaya(ref:8) that predicts domain boundaries by finding regions having small side chain entropy, is performed.

b) Model selection: We first selected structures with energy lower than a cutoff and with the contact order higher than a cutoff from ensemble obtained by multi-canonical ensemble FA simulations. The resulting structures are then clustered. If whole-length structures are not well clustered, the substructures are clustered. The representatives of larger clusters are chosen as models based on human inspection.

c) Models for targets that are likely to be CM & FR: We combine PDB-BLAST or 3D-Jury's alignment with simple loop insertion to make candidates, which is followed by visual inspection to choose the final models.

*Both of them equally contributed to the work.


1. Takada,S. (2001) Protein Folding Simulation With Solvent-Induced Force Field: Folding Pathway Ensemble of Three-Helix-Bundle. *Proteins* **42**, 85-98.
2. Fujitsuka,Y., Takada,S., Luthey-Schulten,Z.A., and Wolynes,P.G. (2004) Optimizing Physical Energy Functions for Protein Folding. *Proteins* **54**, 88-103.
3. Chikenji,G., Fujitsuka,Y., and Takada,S. (2003) A reversible fragment assembly method for de novo protein structure prediction. *J.Chem.Phys*. **119**, 6895-6903.
4. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. **25** 3389-3402.
5. Ginalski,K., Elofsson,A., Fischer,D., Rychlewski,L. (2003) 3D-Jury: a simple approach to improve protein structure predictions, *Bioinformatics* **22**, 1015-1018.
6. Marchler-Bauer,A., et al, (2003) CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res*. **31**, 383-387.
7. Marchler-Bauer,A., et al (2003) CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res*. **30**, 281-283.
8. Galzitskaya,O.V. and Melnik,B.S., (2003) Prediction of protien domain boundaries from seaquence alone. *Protein Science* **12**, 696-701.

**Rokky** (serv) - 279 models for 63 3D targets

## Structure prediction server that integrates PDB-BLAST, 3D-Jury, and the SimFold fragment assembly simulator

W. Jin[1], T. Furuta[1], S.J. Park[2], N. Koga[3], Y. Fujitsuka[3], G. Chikenji[3], and S. Takada[1,3]

*[1] – Faculty of Sci, Kobe Univ, [2] – Interdisciplinary Grad School of Sci & Eng, Tokyo Inst Tech, [3]- Grad School, Sci & Tech Kobe Univ*

stakada@kobe-u.ac.jp

The server Rokky performs the fragment assembly simulated annealing with SimFold energy function for parts of the query sequence that are likely to be new fold, whereas other parts are modeled by either PDB-BLAST or 3D-Jury with variable loops constructed from a library. Individually modeled domains are, if possible, docked to have models of the whole sequence. The Rokky is still premature, has very much evolved through the CASP6 summer and so methods used are somewhat different for targets submitted in different weeks.
Here, we briefly describe 1) job flow, 2) loop modeling, 3) generation of fragment candidates, and 4) fragment assembly with SimFold and model selection. The energy function, SimFold is described in the method of corresponding human prediction team Rokko.

1) Job flow: Rokky is a server which predicts protein 3D-structures automatically. For all targets, the Rokky first performs PSI-BLAST and PDB-BLAST using nr and PDB databases, respectively. When the template with e-value smaller than 0.001 is found in PDB-BLAST, the Rokky uses its alignment and makes model structures by inserting loop structures in alignment gaps. Otherwise, the Rokky submits the target sequence to 3D-Jury meta server and obtains the results. When 3D-Jury-score higher than 30.0 is found, the Rokky uses 3D-Jury's template and its alignment and makes model structures after loop insertion in alignment gaps. For the rest, the Rokky performs the fragment assembly simulated annealing with SimFold energy function for parts of the unaligned sequence and choose 5 models in sampled structures based on clustering analysis.

2) Loop modeling: Generic loop library was constructed for n-residues (4=<n=<30) that are sorted by the end-to-end distance. For each gapped loop in the query sequence, two-residues outside the gapped region in both ends (4 residues in total) are best-fitted to all members in the corresponding part of the library and the loop that has minimal RMSD in these 4 residues are used as a

loop model.

3) Generation of fragment candidates: For every 10-residue in the query sequence, the correlation coefficient of 20×10 dimensional fragment vectors made of the PSSM from PSI-BLAST retrieves 10-residue fragment candidates from 2100 template proteins that have known structures. The collection of 20 fragment candidates for each site of the target overlapping is filtered by Ramachandran plot if PSI-PRED says the site is confidential helix.

4) Fragment assembly (FA) with SimFold and model selection: The server Rokky performs the FA simulated annealing simulation with SimFold using fragment candidates generated by 3) for the targets that has no apparent template. A randomly chosen fragment with the length of 4 aa to 9 aa are replaced with another fragment randomly chosen from the candidate list by the Metropolis judgement at each step. Selection temperature is gradually decreased to obtain low-energy structures. This FA simulated annealing runs are repeated as many samples as possible till a few hours to the deadline (48 hours). The sampled structures that have secondary structure more than a cutoff are treated by the cluster analysis with the group average method, in which centers of the five largest clusters are chosen as final models.


1. Takada,S. (2001) Protein Folding Simulation With Solvent-Induced Force Field: Folding Pathway Ensemble of Three-Helix-Bundle. *Proteins* **42**, 85-98.
2. Fujitsuka,Y., Takada,S., Luthey-Schulten,Z.A., and Wolynes,P.G. (2004) Optimizing Physical Energy Functions for Protein Folding. *Proteins* **54**, 88-103.
3. Chikenji,G., Fujitsuka,Y., and Takada,S. (2003) A reversible fragment assembly method for de novo protein structure prediction. *J.Chem.Phys*. **119**, 6895-6903.
4. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. **25** 3389-3402.
5. Ginalski,K., Elofsson,A., Fischer,D., Rychlewski,L. (2003) 3D-Jury: a simple approach to improve protein structure predictions, *Bioinformatics* **22**, 1015-1018.
6. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195-202


**Rost** - 54 models for 33 3D / 11 FN targets

## A comprehensive manual assessment approach for structure and function predictions of fold recognition CASP6 targets


Kazimierz O. Wrzeszczynski[2], Avner Schlessinger[1], Yana Bromberg[3], Marco Punta[1] and Burkhard Rost[1]

*[1] Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY, [2] Integrated Program in Cellular, Molecular, and Biophysical Studies, Columbia University, New York, NY, [3] Department of Biomedical Informatics, Columbia University, New York, NY.*

rost@columbia.edu


The approach utilized multiple methods in the manual selection process toward a target to template alignment prediction for fold recognition (or difficult comparative modeling - CM/FR) CASP6 targets. In order to classify targets as fold recognition or difficult comparative modeling we mainly relied on two criteria, 1) low sequence conservation (PSI-BLAST[1] E-value >10e-3 and 35% sequence identity) to current PDB[2] templates and 2) all identified PDB templates comprised a large variety of folds as classified by SCOP[3]. The initial template selection for each CASP6 target was performed using PSI-BLAST alignment and the CAFASP (http://www.cs.bgi.ac.il/~dfischer/CAFASP4/) PDB-Blast output. Additional template identification was approached via a multi-step process applying many of the publicly available sequence and structure alignment tools in particular AGAPE[4], 3D-Jury[5], FFAS03[6], along with the secondary structure prediction method ProfSec[7] and two function domain-identifying servers Pfam[8] and CHOP[9]. Manual inspection through visualization of each alignment was then performed for the selection of the final target to template prediction.

We primarily considered results from our in-house alignment method AGAPE (note that this method has been publicly available since the beginning of CASP6). High consideration was also given to the 3D-Jury method or to the various automated fold recognition servers found through CAFASP, specifically those that were able to identify a template structure that encompassed a majority of the target sequence. Finally, a six iteration PSI-BLAST alignment to PDB entries was also considered to identify structural templates when significant structural or functional motifs were aligned between the CASP6 target and the PDB entry. All possible templates were then submitted to AGAPE and FFAS03 for additional pairwise alignment of the target and template. Once several templates were under consideration each initial

alignment (AGAPE, PSI-BLAST or other fold recognition sever) was further scrutinized through manual inspection of the aligned predicted secondary structure (using ProfSec for the target) with each template. Pfam domain alignment was used to identify family conserved residues. An extensive literature search for functional (such as catalytic, metal or ligand binding) or structural important residues (such as disulfide or salt bridges) for each template was performed and greater template preference was given when those corresponding residues correctly aligned with the target. Further manual visual inspection using GRASP2[10] and VMD[11] was performed to check for any inconsistencies in biophysical properties (i.e. exposed hydrophobic residues, hydrogen bond distances, etc.) for each possible model or alignment. Each target to template alignment was then manually adjusted to fit the above given parameters for the final prediction submission.

Function prediction was performed using a combination of methods encompassing domain function identification, motif searches, homology detection, literature searches and catalytic residue alignment. An initial, general function assignment was made through domain detection using CHOP and Pfam and compared with homologues found through PSI-BLAST alignment often identifying the predicted Gene Ontology[12] categories. More intricate function prediction whenever possible was developed mainly through annotation transfer from literature search for functional residues in the identified specific structural template as aligned with the target sequence.


1. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. **25**, 3389-3402.
2. Berman,H.M., Westbrook,J., Feng,Z., Gilland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N., & Bourne,P.E. (2000). The Protein Data Bank. *Nucleic Acids Res*. **28**, 235-242.
3. Murzin,A.G., Brenner,S.E., Hubbard,T., & Chothia,C. (1995). SCOP: a structural classification of proteins database for investigation of sequences and structures. *J. Mol. Biol*. **247**, 536-540.
4. Przybylski,D., Rost,B. (2004). Improving fold recognition without folds. *J. Mol. Biol.* **341**, 255-269.
5. Ginalski,K., Rychlewski,L. (2003). Detection of reliable and unexpected protein fold predictions using 3D-Jury. *Nucleic Acids Res*. **31**, 3291-3292.
6. Rychlweski,L., Jaroszewski,L., Li,W., & Godzik, A. (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci*. **9**, 232-241.
7. Rost, B. (1996) PHD: predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol*. **266**, 525-39.
8. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall, M., Moxon,S., Sonnhammer,E.L., Studholme,D.J., Yeats,C., & Eddy,S.R. (2004) The Pfam protein family database. *Nucleic Acids Res*. **32**, D132-D141.
9. Liu,J., Rost,B. (2004) CHOP: parsing proteins into structural domains. *Nucleic Acids Res*. **32**, W569-W571.
10. Petrey,D., Honig,B. (2003). GRASP2: vizualization, surface properties and electrostatics of macromolecular structures and sequences. *Methods Enzymol.* **374**, 492-509.
11. Humphrey,W., Dalke, A., & Schulten, K. (1996). VMD - Visual Molecular Dynamics. *J. Mol. Graph*. **14**, 33-38.
12. Harris,M.A., Clark,J., Ireland,A., Lomax,J., etc. al; Gene Ontology Consortium. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*. **32**, D258-D261.


## **rost_PROFcon** (serv) - 64 models for 64 RR targets

### PROFcon - a new neural network-based contact predictor


M. Punta[1,2] and B. Rost[1,2,3]

*[1] -CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY 10032, USA, [2] - Columbia University Center for Computational Biology and Bioinformatics (C2B2), New York, NY 10032, USA,, [3] - NorthEast Structural Genomics Consortium (NESG), Department of Biochemistry and Molecular Biophysics, Columbia University, New York, USA*

punta@cubic.bioc.columbia.edu


We introduce a novel contact prediction method (named PROFcon) that combines information from alignments, one-dimensional predictions, from the region between two contacting residues, and the average properties of the entire protein chain. The method is based on a simple feed-forward back-propagation neural network (NN). We train the NN on a large number of proteins (748) and validate the method's performance on sets that differ in protein length, number of aligned homologous sequences, and structural class. PROFcon performance is rather robust as a function of protein length and decreases in the absence of a proper number of aligned homologous sequences (sparse evolutionary profiles). The best accuracy is achieved for proteins belonging to the alpha/beta SCOP[1-2] structural class. In the following we give a more detailed description of dataset selection and of the features used as input to the NN.

*Data sets and cross-validation.* The EVA server evaluating structure prediction methods[3] maintains a continuously updated subset of sequence-unique PDB chains (no pair of proteins in this set has HSSP-value above 0[4-5]). In particular, we use the December 2003 EVA release, a set of 3201 protein chains of known structure.
From this initial list we remove all non-X-ray structures, all membrane and coiled-coil proteins and proteins with physical chain breaks[6]. Then, we divide the X-ray-solved protein list into three sets. For the purposes of training, we select structures with resolution ≤2.0 Å. For the validation process (i.e. optimization of all NN parameters) we use structures with resolution in the interval 2.5-3.0 Å and finally, for testing, structures in the interval 2.0-2.5 Å. Due to computational limitations, we reduce the test set to include only proteins of length maximum 400 aa. Training, validation and test sets contain 748, 466 and 633 proteins, respectively.

*Definition of contact.* Two aa are considered two be in contact if their $C_\beta$ atoms - $C_\alpha$ for glycines – are closer than 8 Å.

*NN architecture overview.* We train a standard feed-forward NN with back-propagation and momentum term[7]. We address the extremely unequal distribution of true (contact) and false (non-contact) by balanced training[7]. Since the NN 'sees' the symmetric pairs *ij* and *ji* as two different samples, the actual PROFcon output value for the *ij* pair is obtained as the average over the *ij* and *ji* NN output[8]. The NN uses 738 input, 100 hidden, and 2 output nodes (contact, non-contact).

*Detailed specification of input.* The pairs are characterized through: 1) local information, 2) connecting segment information, 3) protein information.
*1) Local information: ij centered windows and pair-specific features.* For each residue pair *ij* in a protein, the network incorporates information from aa comprised in two windows of size 9 centered around *i* and *j* (corresponding to intervals [*i*-4;*i*+4] and [*j*-4;*j*+4]). Each sequence position within the two windows is characterized by 29 nodes: 20 for the evolutionary profile (i.e. frequency of occurrence of the 20 aa types at that position, as obtained from multiple sequence alignments[9-10]), one additional node to account for the N and C terminal residues[7], 4 for the predicted secondary structure (three values per residue for helix-strand-other + one value for prediction reliability), 3 for the predicted solvent accessibility (two values for buried-exposed + one value for prediction reliability) and, finally, 1 for the conservation weight[10]. Alignments are obtained through PSI-BLAST[11] filtering the aligned sequences at 80% sequence identity (i.e. any two sequences in the multiple sequence alignment have <80% sequence identity). We predict secondary structure and solvent accessibility using PROFphd[12]. Note that we train and test on predicted rather than observed 1D values. As the two windows together account for 18 positions, we need a total of 522 nodes for their description. Two more features are introduced to better characterize the central residues *i* and *j*. These are: pair type (hydrophobic-hydrophobic, polar-polar, charged-polar, opposite charges, same charges, aromatic-aromatic, other[13]) (7 nodes) and pair complexity (whether or not the two residues are in a low-complexity region, according to SEG[14] (2 nodes).
*2) Connecting segment information: central window, length and average properties.* The segment's central positions have been shown to be the most informative for contacts[6]. So, we introduce a window of size 5 spanning the interval [int(|*i*-*j*|/2)-2; int(|*i*-*j*|/2)+2]. Sequence positions within this window are characterized in the same exact way as positions in the *ij*-centered windows (i.e. 29 nodes each). Further, we use 11 nodes for segment length description, corresponding to sequence separations 6, 7, 8, 9 and to intervals 10-14, 15-19, 20-24, 25-29, 30-39, 40-49, >49 (values chosen by intuition not by optimization). Note that the encoding of segment length was necessary in order to qualitatively reproduce the observed distribution of contact probability versus sequence separation (the shorter the sequence separation, the higher the probability of being in contact)[15]. Finally, we add in nodes encoding for segment's average properties: 20 nodes for aa composition, 3 nodes for secondary structure composition and one node for the fraction of aa in the segment in a low-complexity region. Overall, we use 180 nodes for the description of the segment.
*3) Protein information: length and average properties.* We use 20+3 nodes for the average aa and secondary structure composition of the entire protein, plus 4 nodes to describe the protein length (intervals 1-61, 61-120, 121-240, >241; again, values are chosen by intuition).


1. Murzin,A.G., Brenner,S.E., Hubbard,T.J.., & Chothia C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* **247,** 536-40.
2. Andreeva, A., Howorth,D., Brebber,S.E., Hubbard,T.J., Chothia,C. & Murzin,A.G. (2004). SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res* **32** Database issue: D226-9.
3. Koh,I.Y.Y., Eyrich,V.A., Marti-Renom,M.A., Przybylski,D., Madhusudhan,M.S., Eswar,N., Grana,O., Pazos,F., Valencia,A., Sali,A. & Rost,B. (2003). EVA: evaluation of protein structure prediction servers. *Nucleic Acids Res* **31**(13), 3311-3315
4. Rost,B. (1999). Twilight zone of protein sequence alignments. *Protein Eng* **12**(2), 85-94.

5. Sander,C. & Schneider,R. (1991). Database of homology-derived structures and the structural meaning of sequence alignment. *Proteins* **9,** 56-68.
6. Gorodkin,J., Lund,O., Andersen,C.A. & Brunak,S. (1999). Using sequence motifs for enhanced neural network prediction of protein distance constraints. *Ismb*, 95-105.
7. Rost,B., & Sander,C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol* **232,** 584-599.
8. Pollastri,G. & Baldi,P. (2002). Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners. *Bioinformatics* **18**(Suppl 1), S62-S70.
9. Przybylski,D. & Rost,B. (2002). Alignments grow, secondary structure prediction improves. *Proteins* **46,** 195-205.
10. Rost,B. (1996). PHD: predicting one-dimensional protein structure by profile based neural networks. *Methods Enzymol* **266,** 525-539.
11. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
12. Rost,B. (2004). How to use protein 1D structure predicted by PROFphd. *Methods Mol Biol*: submitted.
13. Creighton,T. (1992). *Proteins: Structures and Molecular Properties*.
14. Wootton,J.C. & Federhen,S. (1996). Analysis of compositionally biased regions in sequence databases. *Methods Enzymol* **266,** 554-71.
15. Fariselli,P. & Casadio,R. (1999). A neural network based predictor of residue contacts in proteins. *Protein Eng* **12**(1), 15-21.

## **SAM-T04-hand** - 375 models for 64 3D / 56 RR targets

### Merging fold-recognition, new-fold and comparative modeling methods

K. Karplus, S. Katzmann, G. Shackelford, M. Koeva, J. Draper, B. Barnes, M. Soriano, R. Hughey
*University of California, Santa Cruz*
karplus@soe.ucsc.edu

The SAM-T04 human predictions for CASP6 use a very similar fold-recognition method to the SAM-T02 method in CASP5 [1].

We start with a fully automated method:

- Use the SAM-T2K and SAM-T04 methods for finding homologs of the target and aligning them.
- Make local structure predictions using neural nets and the multiple alignments. Different neural nets are used for the SAM-T2K alignments and the SAM-T04 alignments. We currently use 7 local-structure alphabets:
  - DSSP
  - STRIDE
  - STR2 - an extended version of DSSP that splits the beta strands into multiple classes (parallel/antiparallel/mixed,edge/center)
  - ALPHA - a discretization of the alpha torsion angle: $C_\alpha$(i-1), $C_\alpha$(i), $C_\alpha$(i+1), $C_\alpha$(i+2)
  - BYS - a discretization of Ramachandran plots, due to Bystroff
  - CB_burial_14_7 - a 7-state discretization of the number of $C_\beta$ atoms in a 14 Angstrom radius sphere around the $C_\beta$
  - DSSP_EHL2 - CASP's collapse of the DSSP alphabet. DSSP_EHL2 is not predicted directly by a neural net, but is computed as a weighted average of the other backbone alphabet predictions.
- We make 2-track HMMs with each alphabet (1.0 amino acid + 0.3 local structure) and use them to score a template library of 6400 (T04) or 9900 (T2K) templates. We also used a single-track HMM to score not just the template library, but a non-redundant copy of the entire PDB.
- We also made a few 3-track HMMs (AA, STR2, CB_burial_14_7) for finding and aligning more remote homologs.
- One-track HMMs built from the template library multiple alignments were used to score the target sequence (for early targets, only T2K template library was searched this way).
- All the logs of e-values were combined in a weighted average (with rather arbitrary weights, since we did not have time to optimize them), and the best templates ranked. Ranking was separate for predictions from the T2K and T04 multiple alignments.
- Alignments of the target to the top templates were made using several different alignment methods (mainly using the SAM hmmscore program, but a few alignments were made with Bob Edgar's MUSCLE profile-profile aligner).
- Generate fragments (short 9-residue alignments for each position) using SAM's "fragfinder" program and the 3-track HMM.
- Then the "undertaker" program (named because it optimizes burial) is used to try to combine the alignments and the fragments into a consistent 3D model. No single alignment or parent template was used, though in many

cases one had much more influence than the others. The alignment scores were not passed to undertaker, but were used only to pick the set of alignments and fragments that undertaker would see.

After the initial automatic run was finished, the results were examined by hand, and various tweaks were made to the undertaker cost function to improve the models. Many of the tweaks consisted of adding specific Hbonds, SSbonds, or distance constraints (often as strand-pairing constraints), to make the model look better to us.

Undertaker uses a genetic algorithm with about 28 different operators to minimize its cost function. The cost function has many components, including various definitions of burial and compactness, sidechain rotamer preferences, steric clashes, chain breaks, predicted local backbone conformation, hydrogen bonding, disulfide bonds, and user specified constraints. The relative weights of these components were tweaked for each target, as we have not found a generally applicable set of weights.

Because undertaker does not (yet) handle multimers, we sometimes added "scaffolding" constraints by hand to try to retain structure in dimerization interfaces, and sometimes did modeling of double-length chains for dimers.

For multiple-domain models, we sometimes broke the sequence into chunks (often somewhat arbitrary overlapping chunks), and did the full method for each subchain. The alignments found were all tossed into the undertaker conformation search. In some cases, we performed undertaker runs for the sub-chains, and cut-and-pasted the pieces into one PDB file (with bad breaks) and let undertaker try to assemble the pieces.

Preliminary analysis of the results indicates that getting a good template and alignment is still overwhelmingly the most important step in getting a good model.

1. Karplus,K., Karchin,R., Draper,J., Casper,J. Mandel-Gutfreund,Y., Diekhans,M., and Hughey,R. (2003) Combining local-structure, fold-recognition, and new-fold methods for protein structure prediction. *Proteins: Structure, Function, and Genetics* **53** (S6), 491-496.

## **SAMUDRALA** - 286 models for 64 3D targets

### Refining comparative models using a graph-theoretic approach

T. Liu and R. Samudrala
*University of Washington*
{tianyun,ram}@compbio.washington.edu

We evaluated the ability and effectiveness of a novel graph-theoretic approach to find the optimal interactions in a protein structure, given a variety of side-chain and main-chain conformational choices for each position. Sampling of side-chain and main-chain conformations was accomplished by exhaustively enumerating all possible choices from a population of initial models. The best combinations of these possibilities were selected through an all-atom scoring function [1] aided by the graph-theoretic approach [2].

For each CASP6 target, several models were generated using 3D-Jury server (http://BioInfo.PL/Meta)[3] combined with our comparative modeling server, PROTINFO-CM (http://protinfo.compbio.washington.edu)[4]. Additional models were obtained from the CAFASP4 server after scrutinizing the alignments to gain extra variability in sequence alignments and templates. Models were inspected for missing or incorrect parts, typically for loops. If reasonable alternative loops could be built using our in-house software, they were added to the pool as well. Side-chain possibilities were also constructed using the program SCWRL[5]. Care was taken to assure that models were superimposed based on their secondary structure so that the average α-carbon root mean square deviation (cRMSD) between each model was less than 5 Å.

After a set of models was superimposed, the next step required the determination of the crossover points where mixing between different parent structures could occur. Crossover points were defined by the ranges of main-chain where the α-carbon was less than 1.0 Å from each other, and were not permitted inside secondary structure elements.

We then used a graph-theoretic clique-finding approach to assemble the sampled side-chain and main-chain conformations. A complete description of the method is given elsewhere[2]. The idea of this approach is to obtain optimized mosaic models by shuffling them in a rational way. Thus the key point is the choice of an appropriate scoring function. We used an all-atom conditional probability discriminatory function (RAPDF)[1] to evaluate the cliques, with the

highest scoring ones representing the optimal combinations of the different side-chain and main-chain possibilities.

In the final step, all models from the above approach were refined with ENCAD[6]. The effectiveness of this methodology to improve the model accuracy remains to be investigated.


1. Samudrala,R., Moult,J. (1998) An all-atom distance dependent conditional probability discriminatory function for protein structure prediction. *J Mol Biol **275***, 893-914.
2. Samudrala,R., Moult,J. (1998) A graph-theoretic algorithm for comparative modelling of protein structure. *J Mol Biol **279***, 287-302.
3. Ginalski,K., Elofsson,A., Fischer,D., Rychlewski,L. (2003) 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics*. ***19***, 1015-1015.
4. Hung,L-H., Samudrala,R. (2003) PROTINFO: Secondary and tertiary protein structure prediction. *Nucleic Acids Research **31***, 3296-3299.
5. Bower,M.J., Cohen,F.E., Dunbrack,R.L. (1997) Prediction of side-chain orientations from a backbone-dependent rotamer library: A new homology modelling tool. *J Mol Biol **267***, 1268-1282.
6. Levitt,M., Hirshberg,M., Sharon,R., Daggett,V. (1995) Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. *Comp Phys Comm **91***, 215-231.


## **SAMUDRALA-AB** - 200 models for 40 3D targets

### Generating, selecting and refining protein structures *de novo*


L-H. Hung, S.C. Ngan, and R. Samudrala
*University of Washington*
{lhhung,ngan,ram}@compbio.washington.edu


We have implemented a new tri-partite protocol for the automated prediction of protein structure from sequence alone. Structures are generated using a simulated annealing search phase that minimizes a target scoring function. Moves are derived from a synthetic function that produces $\phi/\sigma$ angular distributions similar to the empirically observed ones. In contrast to fragment based methods, this is accomplished without copying any angles or coordinates. After the search phase, a local minimization protocol further reduces the target score. In cases where there are strands or constraints, a pre-condensation phase allows strands to pair and constraints to be satisfied. A series of composite functions based on different combinations of 14 individual scoring functions is used to choose a set of best conformers. A novel iterative density protocol is then used to choose the best structures from this set. Finally, the best conformers are used to guide the generation of new conformers, thus iteratively refining the predicted structure. As of this abstract submission, T0236 model 5 is an example of the protocol where a structure better than most of the fold recognition models (cRMSD of 1.97 A for residues 1-50, 5.42 A for all 84 residues) is produced.

Generation of structures is accomplished through a search phase where a composite energy function is minimized by Monte-Carlo simulated annealing. In contrast to methods that replace fragments from known structures, the present protocol uses a function that generates $\phi/\sigma$ angles that reflects the distribution observed in the PDB, and does not copy any angles or coordinates. All residues in a given protein sequence are first classified by the encompassing triplet sequence and the triplet secondary structure. A histogram is then constructed from the $\phi/\sigma$ angles of matching triplets of the same secondary structure in the PDB. (A bin size of 10 degrees by 10 degrees is used and only the angles in the central residue of the triplet are plotted). The mean $\phi/\sigma$ angle in each bin and the standard deviation are recorded. To choose a $\phi/\sigma$ pair during the simulation, a bin is first chosen using the frequencies observed in the histogram. The angles are then chosen using a normal distribution that fits the mean and standard deviation of the observed distribution within the bin.

In addition to the main search phase we have also added a minimization phase using Brent's method and small random moves which typically result in a further 10% reduction in the target score. A pre-condensation phase, implemented late in CASP, encourages pairing of strand residues and satisfaction of other constraints resulting in 10-100 fold increase in the number of paired strands formed.

The search target function is a compromise between the speed of evaluation and the best correlation to the distance from the native structure. We keep the 10 best conformers per seed for analyses using 14 energy-like scoring functions encompassing physical energy functions (*vdw*, *electrostatic*, *solv*), general empirical functions (*Shell*, *MJ*, *hcf*, *Sol*, and *Rad* ) and PDB-based empirical functions (*RAPDF*, *Coord*, *Conseq* and *Curv*). Due to the diversity of both the functions and the proteins that are being evaluated, it is difficult to derive a single weighting scheme that produces an optimal composite function. Instead, the best linear combinations of these functions were determined by logistic regression on large sets of decoys. 19 groups of these linear combinations were

used to filter the initial set of conformers. Typically, 100,000 – 200,000 conformers are reduced to about 1000-2000 at this stage

Energy-like scoring functions alone are still very inconsistent at picking out the best structures. Fortunately, one of the most powerful scoring functions is the completely statistical density function, which is the (negative) sum of RMSD to the other conformations in the set and is a measure of the distance of a conformer to the center of the distribution. Unfortunately, the largest contributions to the density scores come from the outliers that can skew the correlation of density to the distance to the true center of the distribution, reducing the effectiveness of the function. Thus, we have implemented a new iterative density function that measures the density, removes the worst outlier (the conformer with lowest density) and then repeats the process until there are no more outliers in the set. The center of this trimmed set is then selected (and the centers of the largest k-means clusters for the final selection of 5 for CASP) and is taken as the best.

Finally, if there is a good cluster of conformers it is possible to generate a better set of conformers near the conformational center. This is done by incorporating the RMSD to the best conformers into the target function and/or using internal distance constraints derived from these conformers and repeating the generation stage. Selection of the best conformers proceeds as before and the spread of the final set of 5 conformers is reduced to 2-4 A cRMSD.

## **SBC** - 90 models for 64 3D targets

### Use of Pcons, ProQ and Pmodeller in CASP6

Björn Wallner, Tomas Ohlson, Bob McCallum, Arne Elofsson

[1]– *Stockholm Bioinformatics Center*

bjorn@sbc.su.se, tomasoh@sbc.su.se, maccallr@sbc.su.se, arne@sbc.su.se

We have submitted automatic and manual predictions using similar techniques as in CASP5. Predictions have been collected by meta-servers[1], and consensus predictions[2]. We have used the latest versions of the Pcons consensus predictor, Pcons 5, and the homology modeling counterpart, Pmodeller 5. The fifth version of Pcons and Pmodeller have been upgraded since CASP5[3]. The new versions have been designed to be more flexible, by the use of a standardized method to include the performance of individual methods. A backbone only version of ProQ[4] has been included in Pcons5 to evaluate structure quality, while Pmodeller5 uses the old version of ProQ.

There are two difference between Pcons5 and Pmodeller5. Firstly Pcons5 submits the alignment as it is received by the fold recognition server, while Pmodeller5 uses a homology modeling program, MODELLER or nest, to build an all atom model of the targets. Secondly, the scoring function to choose the models differs slightly with a higher emphasis on ProQ in Pmodeller than in Pcons.

Due to the last minute changes in CASP and CAFASP organization we had to use two different meta-servers as the input to Pcons and Pmodeller. The Pcons5 and Pmodeller 5 server predictions are based on the genesilico meta-server and used five fold recognition methods, while prediction based on the bioinfo meta-server, SBC-Pcons5 and SBC-Pmodeller5 (submitted as manual predictions due to the CASP/CAFASP mess), utilized up to twenty different fold recognition servers. These predictions are identical in all aspects except that they have used the results from different fold recognition servers as their input. This gave an unexpected possibility to study the importance on the choice of servers for consensus predictions. In our benchmarks based on earlier CASP and LiveBench[5] results we predict that the difference should be limited.

Our manual predictions have focused on two questions, (i) can we use multiple templates to improve the results from Pcons/Pmodeller and (ii) can we improve the Pcons approach with the use of intermediate sequence searches. To answer these questions we set up a system that allowed us to extract homologs submitted to the meta-server and easily build models full atom models by the use of one or several of the alignments. Models from all servers that provided alignments were automatically built and evaluated using ProQ and other evaluation tools. We have used several different homology modeling programs, but found that for practical use only MODELLER could be used to build models based on multiple templates. Furthermore, we found that the use of multiple templates seemed to improve the models in just a few cases. . The use of intermediate sequence searches provided extra support for the choice of target sequences for several targets.

We have submitted predictions as entries:

1. CASP-SERVER ENTRIES:
   Pcomb (5945-6111-1223 and 1461-7232-1594 for late entries)
   Pcons5-genesilico (7154-1189-4551 and 7082-5331-3841 )
   Pmodeller5-genesilico (8015-1578-6073 and 4916-7057-1687)
2. Automatic submissions in the Manual category
   Pcons5-bioinfo (6533-6220-4531)

Pmodeller5-bioinfo (1391-7191-5375)
3. Manual predictions
SBC (5551-1003-7444)


1. Bujnicki,J.M., Elofsson,A., Fischer,D., Rychlewski,L. (2001) Structure Prediction Meta Server. *Bioinformatics* **17**, 750-751.
2. Lundström,J., Rychlewski,L., Bujnicki,J.M. and Elofsson,A. (2001) Pcons: A neural network based consensus predictor that improves fold recognition. *Protein Science* **10**(11), 2354-2362.
3. Wallner,B., Fang,H. & Elofsson,A. (2003) Automatic consensus based fold recognition using Pcons, ProQ and Pmodeller. *Proteins* **53(S6)**, 534-541.
4. Wallner,B. & Elofsson,A. (2003) Can correct protein models be identified? *Protein Science* **12**(5), 1073-1086.
5. Rychlewski,L., Fischer,D. & Elofsson,A. (2003) LiveBench-6: large-scale automated evaluation of protein structure prediction servers. *Proteins* **53(S6)**, 542-547.


## **Scheraga** - 170 models for 34 3D targets

### Physics-based protein-structure prediction using the UNRES and ECEPP/3 force fields - test on CASP6 targets


S. Ołdziej[1,2], C. Czaplewski[1,2], M. Chinchio[1], M. Nanias[1], J.A. Vila[1,3], M. Khalili[1], Y.A. Arnautova[1], A. Jagielska[1], M. Makowski,[1,2] H.D. Schafroth[1], A. Liwo[1], and H.A. Scheraga[1*]

*[1] – Baker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca, NY, 14853-1301, [2] – Faculty of Chemistry, University of Gdańsk, ul. Sobieskiego 18, 80-952 Gdańsk, Poland, [3] - IMASL-CONICET, Facultad de Ciencias Fisico Matematicas y Naturales, Universidad Nacional de San Luis, Ejercito de los Andes 950, 5700 San Luis, Argentina*

[*]has5@cornell.edu


The structures of the target proteins were predicted for the most part with a hierarchical algorithm consisting of three major stages, in which the tertiary structure is predicted at low resolution and then refined[1,2]. Some of the predictions for the α-helical targets (T0198 and T0221) were carried out using only the all-atom ECEPP/3 force field[3] with surface-solvation models and the electrostatically-driven Monte Carlo (EDMC) method as a search technique[4].

In stage 1 of our hierarchical approach, the protein is represented by a simplified low-resolution united residue (UNRES) model, in which the atoms of the peptide group and side chain of each amino-acid residue are replaced with two centers of interactions: the united peptide group (p) located in the middle between two consecutive α-carbon atoms and the united side chain (SC). The lengths of the virtual $C^{\alpha}\ldots C^{\alpha}$ and $C^{\alpha}\ldots$SC bonds are held fixed, but the virtual-bond angles, the virtual-bond dihedral angles, and the orientations of the $C^{\alpha}\ldots$SC virtual bonds are variable. The interactions of this simplified model are described by the UNRES potential derived from the generalized cumulant expansion of a restricted free energy (RFE) function of polypeptide chains[1]. The cumulant expansion enabled us to determine the functional forms of the multibody terms in UNRES. The individual energy terms were subsequently parameterized by using the quantum-mechanical ab initio energy surfaces of model systems and the potential was fine-tuned by applying our novel hierarchical optimization method targeted at decreasing the energy while increasing the native-likeness of structures of the training proteins[2].

Our conformational space annealing (CSA) method with recent modifications to treat both α– and β-structure[5] was used to search for the lowest-energy families of UNRES conformations. To speed up the search in the case of larger proteins, information from secondary structure prediction by PSIPRED[6] was used in the generation of the initial structures; however, the search was carried out in an unrestricted manner with the UNRES energy function. For very large α-helical proteins, a search with our simplified approach[7] in which α-helices are represented as cylinders was carried out and, for the lowest-energy structures thus obtained, the conformational search was completed with the UNRES force field. For targets T0231 and T0234, as a test, one set of UNRES/CSA searches was started from templates provided by the 3D Jury metaserver[8]; the resulting structures turned out to have low UNRES energies and were, therefore, included among the submitted models.

The five families with the lowest UNRES energy obtained in stage 1 were chosen as models 1-5 and converted to all-atom models in stage 2 by using our energy-based method for the reconstruction of an all-atom polypeptide chain from its $C^{\alpha}$-trace and side-chain-centroid coordinates[9,10]. Finally, in stage 3, the all-atom structures were refined by minimizing their energies with the all-atom ECEPP/3 force field[3] subject to $C^{\alpha}$-distance constraints of the parent UNRES models.


1. Liwo,A. et al. (2004) Parameterization of backbone-electrostatic and multibody contributions to the UNRES force field for protein-structure

prediction from *ab initio* energy surfaces of model systems. *J. Phys. Chem. B.* **108**, 9421-9438.
2. Ołdziej,S. et al. (2004) Optimization of the UNRES force field by hierarchical design of the potential-energy landscape. 3. Use of many proteins in optimization. *J. Phys. Chem. B.* in press.
3. Némethy,G. et al. (1992) Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm with application to proline-containing peptides. *J. Phys. Chem.* **96**, 6472-6484.
4. Ripoll,D.R. et al. (1998) New developments of the electrostatically driven Monte Carlo method: test on the membrane-bound portion of melittin. *Biopolymers* **46**, 117-126.
5. Czaplewski,C. et al. (2004) Improved Conformational Space Annealing method to treat β-structure with the UNRES force-field and to enhance scalability of parallel implementation. *Polymer* **45**, 677-686.
6. McGuffin,L.J. et al. (2000) The PSIPRED protein structure prediction server. *Bioinformatics* **16**, 404-405.
7. Nanias,M. et al. (2003) Packing helices in proteins by global optimization of a potential energy function. *Proc. Natl. Acad. Sci. USA.* **100**, 1706-1710.
8. Ginalski,K. et al. (2003) 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics* **19**, 1015-1018.
9. Kaźmierkiewicz,R. et al. (2002) Energy-based reconstruction of a protein backbone from its α-carbon trace by a Monte Carlo method. *J. Comput. Chem.* **23**, 715-723.
10. Kaźmierkiewicz,R. et al. (2003) Addition of side chains to a known backbone with defined side-chain centroids. *Biophys. Chem.* **100**, 261-280.

## **Schulten-Wolynes** - 20 models for 11 3D targets

### Evolutionary profiles derived from QR factorization of multiple sequence and structural alignments

A. Sethi, J. Eargle, P. O'Donoghue, T. Pogorelov, R. Amaro and Z. Luthey-Schulten
*University of Illinois at Urbana-Champaign*
zan@uiuc.edu

Selection of scaffolds for modeling of the target structures was achieved using a combination of profile searches: a single sequence search against HMMPFAM[1] profiles and a profile search using our evolutionary profiles of the target sequence against the NCBI-NR database using HMMer[2] and BLAST[3] or PSIBLAST[3]. The complete evolutionary profiles of the scaffold proteins are nonredundant profiles composed of sequences and structures selected using our multidimensional sequence QR[4-6] algorithm to efficiently represent the evolutionary space of the sequences and structures in the protein family or superfamily. The evolutionary profiles for the target sequences typically contained proteins with sequence identity > 30%. As the evolutionary profile for the template is composed of a combination of sequences and structures of distant homologs, it allows broader diversity of the sequences. In many cases, the template profile contained sequences from beyond the family to include superfamily members.

The evolutionary profiles for target and template sequences were aligned using profile to profile alignment of CLUSTALW[7]. The secondary structure of the template predicted from PSIPRED[8] was used to improve the alignment of the template profile to that of the scaffold semi-automatically. Three-dimensional models of the target protein were made using the Modeller 6v2[9] package based on the improved alignments. The models were generated using the loop refine routine and constraints were applied on secondary structure, when neccessary. This strategy appears to work best for templates within the same family or related superfamilies as the target.

***1.*** Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffith-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L.L., Studholme,D.J., Yeats,C. & Eddy,S.R. (2004). The Pfam protein families database. *Nucleic Acids Res.* **32**, D138-D141.
***2.*** Durbin,R., Eddy,S., Krogh,A. & Mitchison,G. (1998). Biological Sequence Analysis: Probabilistic models of proteins and nucleic acids. *Cambridge University Press*.
***3.*** Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389-3402.
***4.*** O'Donoghue,P. & Luthey-Schulten, Z. (2003). On the evolution of structure in aminoacyl tRNA-synthetases. *Microbiol. Mol. Biol. Rev.* **67**, 550-573.
***5.*** O'Donoghue,P. & Luthey-Schulten,Z. Evolutionary profiles derived from the QR factorization of multiple structural alignments gives an economy of information. *J. Mol. Biol.* Submitted.
***6.*** Sethi,A., O'Donoghue,P. & Luthey-Schulten,Z. Evolutionary profiles derived from the QR factorization of multiple sequence alignments gives an economy of information. In preparation.
***7.*** Thompson,J.D., Higgins,D.G. & Gibson,T.J. 1994. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment

through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673-4680.
***8.*** Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* ***292***, 195-202.
***9.*** Marti-Renom,M.A., Stuart,A., Fiser,A., Sanchez,R., Melo,F. & Sali,A. (2000). Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 291-325.

## **Shiroganese** - 64 models for 64 3D targets

### Structure prediction with match-node profiles *via* HMMs

M. Sato[1] and K. Horimoto[2]
[1] - *Department of Systems and Information Engineering Graduate School, Maebashi Institute of Technology,* [2] - *Laboratory of Biostatistics, Institute of Medical Science, University of Tokyo*
maki@maebashi-it.ac.jp

We will shortly describe the framework of our prediction method. First, the profile of the target sequence is constructed based on the HMM model. Secondly, the similarity of the target profile is searched against a profile database, and the profile with the highest score is detected. Thirdly, the profile alignment between the target and the highest-score profiles is converted to the corresponding sequence-sequence alignment. Finally, based on the alignment, a 3D model is built by a standard homology modeling method.

One of the most significant features in our method is to adopt a novel profile, referred to as *match-nodes profile*[3]. The match-node profile is constructed via profile of hidden Markov models (HMMs)[1]. The profile of a representative HMM is composed of three types of nodes; match node, delete node, and insert node. The three nodes in the HMM profiles represent the probability distribution of amino acid residues, deletions, and insertions at each site in multiple alignment, respectively. We extract the match-node profile, from the series of probability distributions in the match-nodes, which can describe essential characteristics of the multiple alignment.

To build the HMM, we adopt SAM-T2K[6] with w0.5 script. In the construction of the template profile library by SAM-T2K, we utilize the PDB40D[5] and the non-redundant (NR) database from NCBI[7]. Each sequence in PDB40D is regarded as a template sequence, and its similarity is searched against the NR database from NCBI. The SAM-T2K automatically builds the HMMs corresponding with the template sequences in PDB40D. The HMM of target sequence is also built in the same way. Finally, the match-node profile is extracted from the HMM built by the above procedure; 4289 profiles are stored in the present template profile library

As for the scoring scheme of similarity between the profiles, we adopt the log average scoring based on Bayesian theory[4], which is one of the most suitable for the profile-profile comparison. With the log average scoring, the profiles are locally aligned by the dynamic programming procedure. Using the above procedure, we search the similarity of target profile against the template profile library to detect the profile with the highest score.

To construct the 3D model, the alignment between the target and the highest-score profile is converted to the alignment between the respective sequences through the most probable path on the HMM in the highest-score profile. Thus, we obtain the set of the highest-score sequence, structure, and alignment. Based on the results, MozingerZ (MZ), a homology modeling package.[2], is performed to build a 3D-model of the target sequence.

We would like to thank Dr. Koji Ogata for providing the MozingerZ program and helpful advice to build 3D-models.

1. Karplus,K., Barrett,C., and Hughey,R. (1998) Hidden Markov Models for Detecting Remote Protein Homologies, *Bioinformatics* **14,** 846-856.
2. Ogata,K., Leplae,R., Wodak,S.J., An Energy Based Predictions for Multi-loops of Proteins, in preparation.
3. Sato,M., Sugaya,N., Murakami,H., Imaizumi,A., Aburatani,S., Akutsu,T. and Horimoto,K. (2004) Detection of Remote Homologs Based on Hiddend Markov Model Profile. *Res. Comm. Biochem. Cell Mol. Bio.*, in press.
4. von Ohsen,N., Sommer,I. and Zimmer,R. (2003) Profile-profile alignment: a powerful tool for protein structure prediction, *Pac. Symp. Biocompt*., 252-263.
5. http://astral.stanford.edu/
6. http://www.soe.ucsc.edu/research/compbio/sam.html
7. ftp://ftp.ncbi.nhi.gov/blast/db/nr.Z

**SHORTLE** - 113 models for 53 3D targets

## Homology modeling and new fold prediction by emphasizing local interactions

Q. Fang and D. Shortle
*The Johns Hopkins University School of Medicine, Baltimore, MD USA*
dshortl1@jhmi.edu

The foundation of our approach is modeling the energetics of local side-chain interactions with the peptide backbone and with neighboring side-chains using four statistical potentials based on a common reference state. (1) Psb1 is the propensity of a side-chain at position i to adopt specific values of phi, psi, and chi1[1]. (2) Psb234 is a distance-dependent potential reflecting interactions of the side-chain at i with the six closest peptide-bonds beyond those covered by Psb1 (3) PssR is a distance-dependent potential for interactions of the side-chain at i with side-chains at positions i+1 to i+4 and (4) PssΘ is a dihedral-angle potential between CB-CA(i)-and the subsequent CA-CB bonds at positions i+1 to i+4. All four terms include specification of the secondary structure of residue i as turn, helix, or strand. The reference state is a large ensemble of high resolution crystal structures with side-chains averaged to reflect the frequency of the 20 amino acids in the sequences of the same ensemble of proteins. In effect, all four terms are independent components of the probability P(sequence/structure)[2,3]. Individually each term is estimated to average between -0.15 and -0.4 kcal/mole per residue. In combination they appear to contribute an average of 0.6 kcal/mole of free energy per residue to the stabilization of the native structure relative to the same structure with averaged side chains. Thus correct modeling of these interactions should substantially focus the conformational search to a subspace that includes the native state.

For targets identified by the PSI-BLAST and the bioinfo.pl/meta servers as having a structural homologue with greater than 20% sequence identity, the sequence alignment was inferred from a combination of PSI-BLAST and 3D-Jury output. Segments of the target sequence containing turns/loops plus flanking helices/strands were constructed de novo by recombination of 4-, 5-, and 6-residue pieces of high resolution crystal structures selected for their low local interaction energies. (No use was made of the amino acid sequence of these pieces of protein structure.) When the constructed fragment superposed well with confidently aligned helices/strands in the template, they were saved and later clustered to identify the turn geometry with the highest entropy[4]. In 15% of cases, the alignment was readjusted because of difficulty obtaining good superposition based on the initial sequence alignment.

In the second step, the predicted phi/psi/chi1 angles of turns were included as structural restraints, along with the phi/psi/chi1 angles and CB-CB distances taken from the well-aligned region of the structural homologue. In simulated annealing runs using CNS[5] (version 1.0) in torsion angle dynamics mode, 1000 steps were taken at 2000-10,000 degrees, followed by 1000 steps of cooling to room temperature and a default minimization. With this protocol, between 1000 and 5000 all-atom conformations of the target's three dimensional structure were generated on a small LINUX computer farm. The best 10% of conformations, as scored by sum of z-score of a finely binned all-atom potential and the local potentials described above, were then clustered and for 90% of the targets predicted, the cluster center was submitted as the first model for submission; the conformation with the best all-atom potential was submitted as the second model. In a few cases, when the score of the conformation with best all-atom potential was much lower than that of the cluster center, this lowest energy conformation was submitted as model #1.

Targets lacking readily identifiable structural homologues were tackled as new fold challenges. The secondary structure was predicted by comparing the results of PSIPRED and PROFSEC with three secondary structure profiles generated by threading overlapping pieces of target sequence of length 6, 9, and 12 residues[6], selecting for the 20-30 fragments of native structure with (1) the lowest Psb1 + Psb234 scores, (2) the lowest PssR+PssΘ scores, and (3) the best exposure/burial propensity scores. When the PSIPRED/PROFSEC results disagreed with these profiles, either the secondary structure was left unspecified or that segment of secondary structure was alternately treated as a strand and later as a helix.

Using the predicted secondary structure with some fraction of residues unspecified, fragments of the target corresponding to 2 or 3 segments of helix and/or strand (i.e., including 1 or 2 turns, 25-45 residues in length) were constructed by recombination of overlapping 4-, 5-, or 6-residue pieces of high resolution crystal structures selected for low local interaction energies. For each bump-free recombinant, three non-local energy terms were scored: the radius of gyration, an empirical pair potential energy[7], and a simple statistical potential for the distance and torsion angle defined by the ends of the secondary structures separated by each turn. After setting modest cutoffs for these scores based on a short initial run, approximately 300-2000 fragments were generated for each segment of the target, and then clustered by CB-CB distance matrix error. Nine non-overlapping clusters, each containing of 5% of the ensemble,

were visually inspected, along with the most compact and lowest non-local energy scoring fragments.

Depending on the level of convergence and therefore the confidence in the structure of a given segment, between 1 and 1000 fragments for each of 2 to 4 segments overlapping by a single helix/strand were recombined, with selection for compact, non-bumping conformations with good side-chain pair potential scores. These larger recombinants were scored, clustered, and visually inspected as above, and again, depending on the degree of convergence, a variable number of representative fragments were selected for the next round of recombination. The final structure was either manually assembled from large fragments or manually adjusted from a full-length recombinant, to enforce protein-like compactness and patterns of secondary structure interaction. A final short restrained MD run using CNS was carried out to eliminate steric overlap and restore proper covalent bond lengths and angles at sites of breakage during manipulation.

For most targets, a single global structure (topology) was inferred from inspection of the ensembles of fragments and recombinants. In these cases, only a single model was submitted; submission of additional models with different topologies would amount to educated, but nonetheless unsupported guesses. However, for several targets (primarily all beta proteins), no clear topology could be inferred, so 3 to 5 promising recombinants were submitted without manipulation or refinement by CNS.


1. Shortle,D. (2002) Composites of Local Structural Propensities: Evidence for Local Encoding of Long Range Structure. *Protein Science* **11**, 18-26.
2. Simons,K.T., Ruczinski,I., Kooperberg,C., Fox,B.A., Bystroff,C. and Baker,D. (1999) Improved recognition of native-like protein structures using a combination of sequence dependent and sequence-independent features of proteins. *Proteins* **34**, 82-95.
3. Shortle,D. (2003) Propensities, probabilities, and the Boltzmann hypothesis. *Protein Science* **12**, 1298-1302.
4. Shortle,D., Simons,K.T. and Baker,D. (1998) Clustering of low energy conformations near the native structures of small proteins. *Proc. Natl. Acad. Sci. USA* **95**, 11158-11162.
5. Brunger,A.T., Adams,P.D., Clore,G.M., DeLano,W.L., Gros,P., Grosse-Kunstleve,R.W., Jiang,J.-S., Kuszewski,J., Nilges,N., Pannu,N.S., Read,R.J., Rice,L.M., Simonson,T. and Warren,G.L. (1998). Crystallography and NMR system (CNS): A new software system for macromolecular structure determination, *Acta Cryst.* **D54**, 905-921.
6. Fang,Q. and Shortle,D. (2003) Prediction of Protein Structure by Emphasizing Local Side-Chain / Backbone Interactions in Ensembles of Turn Fragments. *Proteins* **53**, 486-490.
7. Bryant,S.H. and Lawrence,C.E. (1993) An empirical energy function for threading protein sequence through the folding motif. *Proteins* **16**, 92-112.


## **Softberry** - 122 models for 63 3D / 59 DR targets

### SoftPM: Softberry tools for protein structure modeling


V. Solovyev [1,2], D. Affonnikov[2], A. Bachinsky[2], I. Titov[2], V. Ivanisenko[2] and Y. Vorobjev[2]

*[1]- Department of Computer Science, Royal Holloway, University of London, Egham, Surrey TW20 0EX,UK;*

*[2]-Softberry Inc., 116 Radio Circle, Suite 400; Mount Kisco, NY 10549, USA*

victor@cs.rhul.ac.uk


We developed a suite of programs *SoftPM* (*Soft*ware for *p*rotein *m*odeling) that was used to analyze CASP6 models. Initial step in 3D modeling is selection of a template structure for a query sequence, or selection of a set of most similar fragments if we study a new fold, and obtaining template-query sequence alignment. This step is performed by *Ffold* program. *Ffold* alignment is made taking into account sequence similarity, secondary structures of both query and template protein, and solvent accessibility of a template protein. Secondary structure of a query protein is predicted by *PSSFinder* program. Secondary structure and accessibility for a template is calculated by *SSENVID* program. As a result, a set of aligned template-query sequence pairs is obtained. Each alignment generates a model structure, and usually up to 2-4 template-query pairs are selected for further modeling.

Building side chain and loop coordinates for a query protein based on a template structure and sequence alignment is performed by *Getatoms* program. To generate a set of side chain conformations for side chain structure prediction, the program uses backbone-independent rotamer library. Rotamers for each residue are ranked according to their frequency of occurrence (statistical potential) and energy of interaction with backbone (VDW scoring potential[1]). Unfavorable conformations are then filtered out using several single-residue criteria, pairwise VDW interaction energy, and Goldstein DEE algorithm[2]. For remaining rotamers, an optimization procedure is performed to obtain a conformation with minimal VDW energy. The loop modeling

procedure in Getatoms program is as follows. A large set of loop main chain conformations satisfying geometrical loop closure criteria is generated and ranked according their sterical energy of interaction with other parts of protein molecule. Top set of the conformations is subjected to the side chain optimization procedure as described above. A conformation with minimal energy is selected as loop model. This procedure is applied consequently for all the loops modeled.

Models output by Getatoms program are further refined by *Hmod3dMM* program, which performs energy minimization using AMBER force field[3,4]. Hmod3dMM consists of two modules. The first module prepares a molecule topology file, which is then used as an input for molecular mechanical minimization module. Energy minimization is first performed in vacuum, and afterwards the resultant structure is further minimized in water. To handle water-water solvent interactions, Hmod3dMM employs special routines that are considerably faster than the standard ones.

At the final stage, all models are evaluated by *Hmod3Dmd* program, which performs general MD simulation of a protein model structure in an implicit solvent via the simulated annealing protocol in the NTE or NTV ensemble.

In the absence of significant homology with known protein structures the structure of query protein is modeled using the *Cover3D* and *Abini3D* programs. *Cover3D* uses *Ffold* results to cover a query sequence with short similar protein fragments with known 3D structure. It outputs several variants of such coverage, which are used by *Abini3D* to compute a putative 3D model of target sequence. *Abini3D* finds optimal conformation of a set of 3D-fragments representing a target sequence. First, it removes the disordered regions from the coverage and generates a set of distinctive partially compact conformations, which are then optimized by genetic algorithm using simplified model of amino acid residues. Then, the algorithm optimizes the energy function derived from statistics on known tertiary structures. Finally, *Abini3D* restores loop structures and outputs the atomic coordinates of optimal conformation. Resulting models are subjected for further refinement using *Hmod3dMM* program. Then all the models are ranked according to *Hmod3Dmd* criteria. The top ranked model is selected as a final solution.


1. Northrup,S.H., Pear,M.R., Morgan,J.D., McCammon,J.A., Karplus,M. (1981) Molecular dynamics of ferrocytochrome c. Magnitude and anisotropy of atomic displacements. *J. Mol. Biol*. **153**, 1087-1109.
2. Goldstein,R.F. (1994) Efficient rotamer elimination applied to protein side-chains and related spin glasses. *Biophys J*. **66**, 1335-1340.
3. Allen,M.P., Tildesley,D.J. Computer Simulations of Liquids (1987) Oxford University Press, Oxford.
4. Weiner,S.J., Kollman,P.A., Nguyen,D.T., Case,D.A. (1986) An All Atom Force Field for Simulations of Proteins and Nucleic Acids. *J. Comput. Chem*. **7**, 230-252.


## Prediction of intrinsic disordered regions in protein sequences


A.G. Bachinsky[2,3] and V.V. Solovyev[1,2]

*[1]-Department of Computer Science, Royal Holloway, University of London, Egham, Surrey TW20 0EX,UK; [2]- Softberry Inc., 116 Radio Circle, Suite 400 Mount Kisco, NY 10549, USA; [3] - SRC VB "Vector", Koltsovo, Novosibirsk region, 630558, Russia*

victor@cs.rhul.ac.uk


Identification of disordered regions in proteins is very important for their structure prediction and functional characterization, as many intrinsically unstructured protein regions play key roles in cell signaling, regulation and cancer. Here we present a new algorithm of identification of disordered regions in proteins, *Pdisorder*.

Training data.

All disordered data used here (649 sequences containing 61237 disordered positions) were downloaded from http://disorder.chem.wsu.edu. Ordered data were prepared as follows: fragments with accurately determined 3D structures were selected from non-redundant set of PDB sequences (2,017 fragments containing 309,454 ordered aminoacid residues). Propensity of each amino acid residue to be in a disordered regions was calculated as $Q_d(i)=(P_d(i)- P_o(i))/(P_d(i)+P_o(i))$, where $P_d(i)$ and $P_o(i)$, are frequencies for i-th type of aminiacid to be in disordered or ordered regions. Promising composition-based attributes[1] as well as a number of property-based attributes including all properties of AAindex (http://www.genome.ad.jp) were tested together with the propensities on their significance for discrimination of disordered regions.

Recognition procedure.

After testing many approaches, a combination of neural network (NN), linear discriminant function (LDF) and a smoothing procedure was selected for recognition of disordered and ordered residues in proteins. At the first stage, we compute features in a sliding window of 31 residues for neural network (21 inputs, 3 layers of neurons, 10 neurons in each layer, 1 output) and for the

linear discriminant function (46 inputs). The $V_i=(L_i+I_i+R_i)/3$ value is used for determining if the i-th position belongs to an ordered region; where $L_i$, $I_i$, $R_i$ are outputs of the neural net for windows starting at positions i, i-15 and i-30. If the value is very high or very low, the position is assumed to be "possibly ordered" – 'o' or "possibly disordered" – 'd'. For intermediate values ('p'), an attempt to use the same outputs of LDF to improve assignment is made. At the second stage, we apply a smoothing procedure that computes chances for the positions of query sequence to be in ordered regions. Assignment of short sequences of "possibly disordered", "possibly ordered" or "unknown" positions is changed depending on sequences surrounding them. The main result of the procedure is a set of long uniform regions with minimum presentation of unknown positions.

Table 1. Estimates of accuracies of different approaches.

| Discriminator | DIS | ORD | AVER |
|---|---|---|---|
| NN | 73.1 | 93.9 | 90.5 |
| LDA | 72.4 | 85.9 | 83.7 |
| NN+LDA | 76.3 | 90.6 | 88.2 |
| **NN (+LDA)** | **83.4** | **93.8** | **92.1** |
| LDA (+NN) | 72.4 | 85.9 | 83.7 |

DIS: Disordered positions, ORD: Ordered positions, AVER: Average accuracy. NN: Only neural net is used, LDA: Only linear discriminant is used, NN+LDA: average values of both outputs are used, NN (+LDA): LDF outputs are used for correction of "unknown" assignments of NN outputs, LDA (+NN): NN outputs are used for correction of "unknown" assignments derived of LDF outputs.

The accuracy of our disorder regions predictor Pdisorder on several test sets is higher than that for the other disorder fragments identification programs such as PONDR[2] and GlobPlot[3].


1. Li,X., Obradovic,Z., Brown,C., Garner,E., Dunker,A. (2000) Comparing predictors of disordered protein. *Genome Informatics* **11**,172-184.
2. Li,X., Romero,P.M., Rani,M.A., Dunker,A., Obradovic,Z. (1999) Predicting protein disorder for N-, C-, and internal regions. *Genome Informatics* **10,** 30-40.
3. Linding,R., Russell,R.B., Neduva,V. and Gibson,T.J. (2003) GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Research* **31**(13), 3701-3708.


## **SSEP-Align** (serv) - 501 models for 63 3D / 61 DP targets

### SSEP – secondary structure elements and profiles


J.E. Gewehr, A.R. Macri and R. Zimmer

*Practical Informatics and Bioinformatics Group, Institut für Informatik,*

*Ludwig-Maximilians-Universität München,*

*Amalienstr. 17, D-80333 Munich, Germany*

{jan.gewehr, alessandro.macri, ralf.zimmer}@bio.ifi.lmu.de


Material

Our template library of domains with known structure is a subset of the ASTRAL compendium[1] The subset which was filtered for 95% sequence identity (ASTRAL95) can be obtained from the ASTRAL website (http://astral.berkeley.edu/). For generating profiles we used the NR database (http://www.ncbi.nlm.nih.gov/) and PSIPRED[2] which relies on PSI-BLAST[3].

Pattern searches were performed with InterProScan on the InterPro database collection[4]. We utilized the MaxSprout[5] server for postprocessing of our C-alpha models to obtain a structure description suitable for CASP submission.

Domain Detection

The SSEP domain prediction server consists of three consecutive steps:

1. Finding Potential Domain Boundaries: Given a target sequence s and a template library of domains, we align each domain sequence with appropriate windows on s with secondary structure element alignment[6]. After discarding insignificant hits, we extract potential domain boundaries from the start and end points of the top-scoring windows.

2. Similarity Scoring of Domain Regions: We define a potential domain region as a subsequence of s that starts and ends with two boundaries. For each region r, we compute a similarity score by aligning each template domain against r, using a combination of secondary structure element alignment and log average profile-profile alignment on both sequence and secondary structure profiles[7]. We also add significant InterPro patterns found on the target s to the set of potential domain regions.

3. Combining Multiple Domain Regions: We rank all valid combinations of non-overlapping domain regions according to a simple combination score based

on the previously computed similarity scores and penalties for unclassified regions. As result we return the five top scoring combinations.

Structure Prediction

The domain prediction is further used to predict the three-dimensional structure. For each potential domain the corresponding segment of the profile is aligned to each ASTRAL profile and the results are ranked according to their score. Depending on the scores and scoregaps between the highest ranking alignments five alternatives are chosen; e.g. in the case of a high score gap to all other domain combinations and high scores for all profile-profile alignments of detected domains within the combination, all five candidates may just be variants of the same alignment. The candidates are then assembled from the corresponding pdb-style-files and stripped down to their C-alpha-coordinates. The full structure for each segment was then reconstructed via MaxSprout.

Whenever the domain detection does not come up with a significant hit and the segments for profile-profile alignment are thus unavailable, we apply once again PSI-BLAST as a means to grab at least partial alignments not covering whole domains. The partial alignments are then used to build models via MaxSprout from C-alpha coordinates.


1. Chandonia,J.M., Hon,G., Walker,N.S., Lo Conte,L., Koehl,P., Levitt,M. & Brenner,S.E (2004). The ASTRAL compendium in 2004. *Nucleic Acids Res.* **32**, D189-D192.
2. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* ***292***, 195-202.
3. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
4. Mulder,J. et al. (2003). The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.* **31**, 315-318.
5. Holm,L. & Sander,C. (1991). Database algorithm for generating protein backbone and side-chain co-ordinates from a C alpha trace application to model building and detection of co-ordinate errors. *J. Mol. Biol.* **218**, 183-194.
6. McGuffin,L.J., Bryson,K. & Jones,D.T. (2001). What are the baselines for protein fold recognition? *Bioinformatics* **17**, 63-72.
7. Öhsen,N.v., Sommer,I., Zimmer,R. & Lengauer,T. (2004). Arby: Automatic Protein Structure Prediction using Profile-Profile Alignment and Confidence Measures. *Bioinformatics* **20**, 2228-2235.


## **Strx_Bix_Geneva** - 15 models for 15 3D targets

### Human intervened comparative modeling in CASP6


Y.L. Yip[1,2] and H. Scheib[1,2]

*[1] – Structural Bioinformatics Group, Department of Structural Biology and Bioinformatics, University of Geneva, Switzerland*

*[2] – Swiss Institute of Bioinformatics*

holger.scheib@isb-sib.ch


Template selection. PSI-BLAST[1] and, if necessary, 3D-PSSM[2] were run with standard parameter settings to identify suitable template structures. In cases where multiple templates were available for a target, the quality of several template structures was assessed from their 'header' entries focusing on i.e. resolution, number of residues in experimental structure, no missing residues or atoms. The Cα atoms of the remaining structures were superimposed in SPDBV[3] to investigate whether templates might exist in more than one conformational state.

Target to template alignment. The target sequence and the selected template structure(s) were aligned in the SPDBV comparative modeling environment. Target to template(s) alignment was guided by ClustalW[4] multiple sequence alignments, secondary structure prediction from both Jpred[5] and Predict Protein servers[6] as well as literature analysis. All alignments were carried out interactively in SPDBV and potential loop anchor points were already identified at this step.

Model building. Models of the structurally conserved regions (all backbone without loops) were built in SPDBV.

Loop building. Loops were generated semi-automatically applying the "Build Loop…" and "Scan Loop Database…" options in SPDBV, respectively. Anchor points were identified from the target to template alignment. Each loop was evaluated according to a simple clash score, overall packing of the protein and biological impact, if one was obvious.

Model refinement and evaluation. Models were refined energetically applying 100 to 200 steps of Steepest Descent using the SPDBV implementation of the GROMOS96 force field[7]. Unfavorable side chain conformations were identified using the "Amino Acids Making Clashes…" and "Amino Acids Making Clashes With Backbone…" options together with a force field energy report.

Problematic backbone conformations were identified applying the Ramachandran plot implemented in SPDBV. We routinely applied SCWRL3.0[8] to optimize side chain placement, but in most cases discarded the results. "Scwrled" and "unscwrled" models were evaluated using Anolea[9], Z-packing by the WhatCheck server[10]. In both cases, the "unscwrled" models yielded better results for most targets.

NOTE: We regarded CASP6 as ideal for teaching and training "on the job".


1. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
2. Kelley,L.A., MacCallum,R.M. & Sternberg,M.J. (2000). Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.* **299**, 499-520.
3. Guex,N. & Peitsch,M.C. (1997). SWISS-MODEL and the Swiss-PDBViewer: an environment for comparative protein modeling. *Electrophoresis* **18**, 2714-2723.
4. Chenna,R., Sugawara,H., Koike,T., Lopez,R., Gibson,T.J., Higgins,D.G. & Thompson,J.D. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* **31**, 3497-3500.
5. Cuff,J.A., Clamp,M.E., Siddiqui,A.S., Finlay,M. & Barton,G.J. (1998). Jpred: a Secondary Structure Prediction Server. *Bioinformatics* **14**, 892-893.
6. Rost,B. (1996) PHD: predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzym.* **266**, 525-539.
7. van Gunsteren,W.F., Billeter,S.R., Eising,A.A., Hünenberger,P.H., Krüger,P., Mark,A.E., Scott,W.R.P. & Tironi,I.G. (1996) Biomolecular Simulation: The GROMOS96 Manual and User Guide, *Vdf Hochschulverlag AG an der ETH Zürich, Zürich Switzerland*, 1-1042.
8. Canutescu,A.A., Shelenkov,A.A. & Dunbrack,R.L. Jr. (2003) A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci.* **12**, 2001-2014.
9. Melo,F. & Feytmans,E. (1997) Novel knowledge-based mean force potential at atomic level. *J. Mol. Biol.* **267**, 207-222.
10. Rodriguez,R., Chinea,G., Lopez,N., Pons,T. & Vriend,G. (1998) Homology modeling model and software evaluation: three related resources. *CABIOS* **14**, 523-528.


## **Taylor** - 193 models for 64 3D targets

### Dynamic domain threading


W.R. Taylor[1,2], T.J. Sheldon[1], K. Lin[1] and I. Jonassen[2]

*1Mathematical Biology,the National Institute for Medical Research, the Ridgeway, Mill Hill, London NW7 1AA, U.K.*

*2 –Biology Unit,University of Bergen,Bergen, Norway.*

wtaylor@nimr.mrc.ac.uk


Each target sequence was scanned across the non-redundant protein sequence databank using PSI-BLAST1 followed by QUEST2 then aligned with MULTAL3. Cutoffs were adjusted until between 10 and 20 maximally disparate sequences remained in the alignment. Secondary structure for each sequence in the alignment was predicted with PSI-PRED4 using just the sequences in the alignment as a mini-database.

To find potential templates, the PSI-BLAST profile was rescanned against the PDB sequences and the target sequence was scanned against a reduced PDB sequence set using TUNE5 and genTHREADER6. All PDB hits were taken as possible templates, plus any that had been identified by QUEST in the original scan. Typically 30 templates were used (10 TUNE, 10 genTHREADER, a maximum of 10 from PSI-BLAST and any others from QUEST).

Complete models for the target were constructed using the program RAMBLE7 which is a simple random-walk based approach previously used to construct 'decoy' proteins then modified to incorporate suitable torsion values in predicted secondary structure segments. In this application, a further constraint is imposed to 'encourage' the selected residue position to lie close to given target points. This is a development from the model construction in the MST8 program used in previous CASPs in which the template positions were retained and 'random' connecting loops (or termini) 'grown' where necessary.

For each template, every buried position in a secondary structure was taken as a starting point from which the chain was grown over the template (using predicted secondary structure where no other guide was available). This was repeated for each predicted secondary structure variant (one for each sequence in the alignment) and repeated five times with different random number seeds. From each starting position, the template was also restricted to a number of residues that matched the target length using an Ising-based domain definition method9. The selected positions were then matched to the target alignment using a scoring scheme similar to that used previously in the MST program.

Depending on the number and size of templates and the number of sequences, the method usually produced between 5,000 to 10,000 models. For clear homologues this gives a dense covering of minor variations whereas for ab-initio prediction (where the selection of templates is effectively random) there is much greater variation. The same method was run on all targets, irrespective of their degree of difficulty.

The resulting alpha-carbon models were ranked using a variety of structure evaluation methods. These included the correlation of conserved hydrophobicity with solvent exposure (as estimated by POPS[10] on the CA positions), the statistical (artificial neural net) method TUNE, the 3D pattern-based method SPREK[11] and the CAO[12] method which is based on correlated amino acid changes in the multiple sequence alignment.

All models were ranked on a combination of these scores and the best assessed visually. For clear homologues, the top 3 ranked models were taken usually without any "user" interference. For the more difficult targets, typically, the top 10 models were considered with some being rejected if they contained any unprotein-like features. The top 3 selections (sometimes more for uncertain results) were then converted to full main-chain models using the CA2MAIN program (Taylor, unpublished) and submitted.


1. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
2. Taylor,W.R. (1998) Dynamic databank searching with templates and multiple alignment. *J. Mol. Biol.* **280**, 375-406.
3. Taylor,W.R. (1988) A flexible method to align large numbers of biological sequences. *J. Mol. Evol.* **28**, 161-169.
4. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195-202.
5. Lin,K., MayA.C. & Taylor,W.R. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *Bioinformatics* **18**, 1350-1357.
6. Jones,D.T (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* **292**, 195-202.
7. Taylor,W.R. and Munro,R.E.J. and Petersen,K. & Bywater,R.P. (2003) Ab initio modelling of the N-terminal domain of the secretin receptors. *Compu. Biol. Chem.* **27**, 103-114.
8. Taylor,W.R. (1997) Multiple sequence threading: an analysis of alignment quality and stability. *J. Mol. Biol.* **269**, 902-943.
9. Taylor, W.R. (1999) Protein structure domain identification. *Prot. Eng.*, **31**, 3364-3366.
10. Cavallo,L., Kleinjung, J. & Fraternali, F. (2003) POPS: A fast algorithm for solvent accessible surface areas at atomic and residue level. *Nucleic Acids Res.* **31**, 3364-3366.
11. Taylor,W.R. & Jonassen,I. (2004) A spattern based method for protein fold recognition. *Proteins* **56**, 222-234.
12. Lin,K., et al. (2003) Testing homology with CAO: a contact-based Markov model of protein evolution. *Comp. Chem.* **27**, 93-102.


## **THGLAB** - 135 models for 27 3D targets

### Coarse-grained protein models and global optimization approaches to protein structure prediction


Matthew S. Lin[1], Brian Carnes[1], Devin Hendricks[1], Nicolas J. Fawzi[1] and Teresa Head-Gordon[1,2]

*1 UCSF/UCB Joint Graduate Group in Bioengineering, 2Department of Bioengineering, University of California, Berkeley*

TLHead-Gordon@lbl.gov


The *ab initio* protein structure prediction problem is to predict the three-dimensional topology of the native state of a protein given its sequence of amino acids. Within an *ab initio* framework, a quantitative description of the free energy surface describing both the proteins' intramolecular forces and the intermolecular interactions with aqueous solvent is required. Because the energy landscape of a realistic-sized protein has thousands of parameters and an enormous number of local minimizers that are potential false traps for the global optimum or very low-lying free energy minimum of the target native structure, global optimization is a promising approach to searching the free energy surface. This year we used two radically different energy functions and global optimization strategies, one that we introduced in CASP5[1-5], and continued to use for the first half of CASP6, and a new strategy based on coarse-grained protein models developed in the Head-Gordon group which were deployed in the second half of the CASP6 competition [6-9].

The first global optimization approach used is our method known as Stochastic Perturbation with Soft Constraints (SPSC)[1-5], which uses Psi-Pred server predictions of secondary structure[10] as mathematical constraints on the optimization search. We do not use a tertiary structure template, or use tertiary structure predictions for generating the terms of the physics-based energy

function[1-5]. The SPSC algorithm is a two-phased approach in which the first phase generates starting structures that are local minima containing predicted secondary structure, and the second phase improves upon the starting structures using both global and local optimizations. All starting structures in the first half of CASP6 were generated with an inverse kinematics (IK) tool developed by Kreylos and co-workers[11], which allows for interactive manipulation of local and global dihedral angle moves, consistent with the Psi-Pred predictions. It was used to form α–helices, as well as an exhaustive enumeration of all possible β–sheet topologies for proteins with predicted strands, including those described in [12], but do not contain any significant tertiary structure. Phase II improves those configurations through global minimizations in a sub-space of the torsion angles of amino acids predicted to be coil. The global optimization produces a number of local minimizers in the subspace chosen, and those conformations are locally minimized in the full variable space. The new minimizers obtained from the local minimizations are merged into the current list, are clustered and ordered by energy value, and the second phase starts again. The process repeats for a number of iterations, until no further progress is made in energy lowering. Local and global optimization algorithms run in parallel on the IBM/SP cluster using up to 512 processors, or across a local cluster of G5s, Alphas, and x86 machines. The parallelization uses a new load balancing technique that is general for large tree search problems using a hierarchical approach[13].

Our description of the protein intramolecular interactions uses the AMBER96 molecular mechanics energy function. For intermolecular interactions, the use of an explicit water potential is computationally expensive in the context of global optimization, and is possibly not needed in structure prediction if the physics of hydration can be adequately described by coarse-grained models. Our research group has studied a critical influence of aqueous solvent on protein conformation, namely hydrophobic interactions, using both experimental solution scattering and simulation[14-16]. The benefits of our hydrophobic hydration function are (1) it is a well-defined model of the hydrophobic effect, (2) it is described by a continuous potential that is more computationally tractable than solvent accessible surface area models, and (3) its novelty in the context of structure prediction of the extra stabilization at a longer length scale for the hydrophobic interaction that is not described by surface area solvation models[14-16].

We have recently tested the model on the publicly available Decoys 'R' Us database[17] and the one created by the Baker group[18] to examine the ability of our energy function to detect the native protein structure from a large set of misfolded or "decoy" structures[19]. We analyzed the performance of our energy function on 20 different proteins (seven α–helical, five β–sheet, and eight α/β proteins), half selected from the Decoys 'R' Us database and the other half from the Baker decoy sets. The potential energy of each protein's native structure was either the lowest or within the lower 5$^{th}$ percentile[19]. The result shows that our energy function can discriminate the native protein structure from a large number of decoy structures.

For targets predicted in the second half of CASP6, we expanded our global optimization strategy to include predictions based on simulated annealing of a coarse-grained protein model developed in the Head-Gordon laboratory[6-9]. For the last 4 targets we relied exclusively on the new approach. The protein chain is modeled as a sequence of beads of three types, hydrophilic, hydrophobic, and neutral, designated by L, B and N, respectively. The pair-wise interaction between beads is attractive for B-B interactions, and repulsive for all other bead pairs (although the strength of the repulsive interactions depend on the bead types involved). In addition to pair-wise non-bonded interactions, the other contributions to the potential energy function include bending and torsional degrees of freedom. Inspired by a simple 2D-model of water[20-22], an additional interaction representing hydrogen bonds between β-strands was added to beads predicted to be in a β-sheet, regardless of bead type[23]. Note that while the non-bonded potential is symmetric with respect to inversion, the dihedral interactions are not symmetric with respect to indice permutations, and we do not find mirror image states.

The general mapping between 20-letter codes provided by CASP and the 3-letter code of the coarse grained model relied on standard interpretations of amino acids as hydrophobic, hydrophilic, or small/neutral. Although there is ambiguity in the mapping from a 20-letter to a 3-letter code, in general the mapping used in Table I of [7] is a good approximation to the original sequence. The model requires assignment of secondary structure, which again is based on the Psi-Pred server. For each target, a maximum of 100 simulated annealing trajectories each with 3 heating/cooling/quench phases were run for a maximum of 24 million timesteps to generate many low energy configurations. These configurations are $C_\alpha$ traces that were then converted to an all-atom model by using CHARMM[24], and a limited memory BFGS local minimization is then used to relieve bad contacts and to improve secondary structure hydrogen-bonding. The structures obtained from the local minimizations are merged into the current list, are clustered and ordered by energy value and examined for good secondary structure content.

Because we chose to focus our efforts on new fold targets, we used BLAST[25-27] and 3D-PSSM[28-30] servers' results to determine which targets met our criteria for probable new folds or sufficiently difficult fold recognitions. We submitted predictions on 30 targets ranging in size from 53 to 417 amino acids.

1. Crivelli,S., Head-Gordon,T., Byrd,R. H., Eskow,E., Schnabel,R. (1999). A hierarchical approach for parallelization of a global optimization method for protein structure prediction. *Lecture Notes in Computer Science, Euro-Par '99*, 578-585.
2. Crivelli,S., Philip,T.M., Byrd,R., Eskow,E., Schnabel,R., Yu,R.C., Head-Gordon,T. (2000). A global optimization strategy for predicting protein tertiary structure: a-helical proteins. *Comp. & Chem.* **24**, 489-497.
3. Azmi,A., Byrd,R. H., Eskow,E., Schnabel,R., Crivelli,S., Philip,T.M., Head-Gordon,T. (2000). Predicting protein tertiary structure using a global optimization algorithm with smoothing. In Optimization in Computational Chemistry and Molecular Biology: Local and Global Approaches, (Kluwer Academic Publishers, Netherlands), 1-18.
4. Crivelli,S., Eskow,E., Bader,B., Lamberti,V., Byrd,R., Schnabel,R., Head-Gordon,T. (2002). A physical approach to protein structure prediction. *Biophys. J.* **82**, 36-49.
5. Eskow,E., Bader,B., Byrd,R., Crivelli,S., Head-Gordon,T., Lamberti,V. Schnabel,R. (2004). An optimization approach to the problem of protein structure prediction. *Math Programming Series A (published online Feb'04)*.
6. Head-Gordon,T., Brown,S. (2003). Minimalist models for protein folding and design. *Curr. Opin. Struct. Biol.* **13**, 160-167.
7. Brown,S., Fawzi,N., Head-Gordon,T. (2003). Coarse-grained sequences for protein folding and design. *Proc. Natl. Acad. Sci* **100**, 10712-10717.
8. Brown,S., Head-Gordon,T. (2004). Intermediates in the folding of proteins L and G. *Prot. Sci.* **13**, 958-970.
9. Sorenson,J.M., Head-Gordon,T. (2002). Toward minimalist models of larger proteins: a ubiquitin-like protein. *Proteins: SFG* **46**, 368-379.
10. McGuffin,L.J., Bryson,K., Jones,D.T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics* **16**, 404-405.
11. Kreylos,O., Hamann,B., Max,N., Bethel,W., Crivelli,S. (2002). Interactive Protein Manipulation, Tech. Report CSE-2002-28, UC Davis.
12. Ruczinski,I., Kooperberg,C., Bonneau,R., Baker,D. (2002). Distributions of beta sheets in proteins with application to structure prediction. *Proteins: SFG* **48**, 85-97.
13. Crivelli,S., Head-Gordon,T. (2004). A new load balancing strategy for the solution of dynamical large tree search problems using a hierarchical approach. *IBM RD Journal* **48**, 153-160.
14. Pertsemlidis,A., Soper,A.K., Sorenson,J.M. Head-Gordon,T. (1999). Evidence for microscopic, long-range hydration forces for a hydrophobic amino acid. *Proc. Natl. Acad. Sci.* **96**, 481-486.
15. Sorenson,J.M., Hura,G., Soper,A.K., Pertsemlidis,A. Head-Gordon,T. (1999). Determining the role of hydration forces in protein folding. Feature Article for *J. Phys. Chem. B* **103**, 5413-5426.
16. Hura,G., Sorenson,J.M., Glaeser,R.M., Head-Gordon,T. (1999). Solution x-ray scattering as a probe of hydration-dependent structuring of aqueous solutions. *Perspectives in Drug Discovery and Design* **17**, 97-118.
17. Samudrala,R., Levitt,M. (2000). Decoys 'R' Us: a database of incorrect onformations to improve protein structure prediction. *Protein Sci.* **9,** 1399.
18. Tsai,J., Bonneau,R., Morozov,A.V., Kuhlman,B., Rohl,C., Baker,D. (2003). An improved protein decoy set for testing energy functions for protein structure prediction. *Proteins: SFG* **52**, 76.
19. Lin,M.S. and Head-Gordon, T.(2004). In preparation.
20. Ben-Naim,A. (1971). Statistical mechanics of "waterlike" particles in two dimensions. I. Physical model and application of the Percus-Yevick equation.*J. Chem. Phys.* **54**, 3682-3695.
21. Ben-Naim,A. (1972). Statistical mechanics of "waterlike" particles in two dimensions. II. One component system..*Mol. Phys*. **24**, 705-721.
22. Truskett,T., Dill,K., (2002) A simple statistical mechanical model of water *J. Phys. Chem. B* **106**, 11829-11842
23. Fawzi,N. J., Yap,E., Head-Gordon,T. (2004). In preparation.
24. Brooks,B.R., Bruccoleri,R.E., Olafson,B.D., States,D.J., Swaminathan,S., Karplus,M. (1983). CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations*J. Comp. Chem.* **4**, 187-217.
25. Altschul,S.F., Gish,W., Miller,W., Myers,E.W., Lipman,D.J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410.
26. Madden,T.L., Tatusov,R.L., Zhang,J. (1996). Applications of network BLAST server. *Meth. Enzymol.* **266**, 131-141.
27. Zhang,Z., Schwartz,S., Wagner.L., Miller,W., (2000). A greedy algorithm for aligning DNA sequences. *J. Comp. Biol.* **7**, 203-214.
28. Kelley,L.A., MacCallum,R.M., Sternberg,M.J.E (2000). Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Bio.* **299**, 499-520.
29. Fischer,D., Barret,C., Bryson,K., Elofsson,A., Godzik,A., Jones,D., Karplus,K.J., Kelley,L.A., MaCallum,R.M., Pawowski,K., Rost,B., Rychlewski,L., Sternberg,M.J (1999). CAFASP-1: critical assessment of fully automated structure prediction methods. *Proteins: SFG*, Suppl **3**, 209-217.
30. Kelley,L.A., MacCallum,R., Sternberg,M.J.E. (1999). Recognition of remote protein homologies using three-dimensional information to generate a position specific scoring matrix in the program 3D-PSSM. *RECOMB99* 218-225.

**TOME** - 453 models for 63 3D targets

## Combined use of @TOME and ViTO at CASP6

V. Catherinot[1] , L. Martin[1], J-L Pons, D. Douguet and G. Labesse[1]
*[1] – Atelier de Bio- et Chimie-informatique Structurale. Centre de Biochimie Structurale, CNRS UMR5048- INSERM U554 - Université Montpellier I ; 34000 Montpellier, France*
labesse@cbs.cnrs.fr

A meta-server, named @TOME (http://abcis.cbs.cnrs.fr/atome), has been previously developed for fold-recognition. It was evaluated during the former CASP experiment (summer 2002) in all the categories ranging from easy comparative modeling to *ab initio* predictions. While both a fully automatic procedure and an expert mode were evaluated at CASP5, we focused, during CASP6, on the expert mode and more specifically on the use of a new software: ViTO[1]. The latter combined an multiple-sequence alignment editor and a 3D visualization tool. This feature facilitates refinement of structural alignments produced by fold-recognition programs. It also allows refined analysis of three-dimensional models produced by SCWRL 3.0[2] or MODELLER 6.2[3] (automatically by @TOME or after manual refinement of the alignments). Furthermore, we attempt to improve theoretical models by variations of some parameters in MODELLER scripts (deviation of models, secondary-structure restraints, …). Similarly, modeling in the context of the quaternary structure was attempted when possible (predicted conservation of the same oligomery for the target and for the template). Best models were chosen according to their scores computed by three evaluation programs: PROSA[4], Verify3D[5] and ERRAT[6].

1. Catherinot,V. & Labesse,G. (2004). ViTO. *Bioinformatics.* In the press.
2. Canutescu,A.A., Shelenkov,A.A. & Dunbrack,R.L.Jr. (2003) A graph-theory algorithm for rapid protein side-chain prediction. *Prot. Sci.*, 1**2**, 2001-2014.
3. Fiser,A. & Sali,A. (2003) Modeller: generation and refinement of homology-based protein structure models. Methods Enzymol **374**,461-91.
4. Sippl,M.J. (1993). Recognition of errors in three-dimensional structures of proteins. *Proteins* **17**, 355-362.
5. Eisenberg,D., Luthy,R. & Bowie,J.U. (1997). VERIFY3D: assessment of protein models with three dimensional profiles. *Methods Enzymol*. **277**, 396-404.
6. Colovos,C. & Yeates,TO. (1993) Verification of protein structures: patterns of nonbonded atomic interactions.. *Prot. Sci.* **2**, 1511-1519.

**UAlbany** - 57 models for 57 DP targets

## Combining neural network and support vector machine classifiers to predict protein domain boundaries

I. MacDonald and G. Berg
*University at Albany, SUNY*
berg@cs.albany.edu

Protein structural domains have become an area of increased interest in structural biology and related disciplines. It is no surprise that techniques to predict aspects of domains, such as the boundaries between them, are emerging. Although exact definitions of domains, boundaries, *etc*. are still emerging[1], work can be done on the basis of existing definitions. For example, the CATH[2] database of protein domain assignments provides labeled structural domain data at the amino acid level. Proteins in the CATH database can be considered as either having one or more domains. A domain boundary is defined as the transition between two different domains in a protein. The focus of our work is the prediction of these domain boundaries in multiple-domain proteins.

Our method is based in the machine learning tradition. Using labeled data, we train a learner on the task of interest. For our data, we collected protein domain information from CATH database entries. The data was screened for quality and manually labeled. Each amino acid in an included protein chain contained as input information a window of that residue and its immediate five neighbors in each direction. The output information was whether or not that residue was part of a boundary region between two structural domains.

The protein chains represented were randomly divided into training and testing sets. Training data was used for the actual training of the predictor. The test data was not used for the training, but rather to gauge when the predictor was performing best on data which it had not previously seen. This provided a simple form of cross-validation in an effort to produce a predictor that was general – that is, not over-fitting to its training data.

In order to increase performance, two separate learners were trained and their combined result was used for predictions. An error-backpropagation artificial neural network and a support vector machine were used as our automated classifiers. Each classifier was independently trained using the training data. The basic neural network used a general back-propagation algorithm[3]. In

situations where the number of instances of one class (non boundary residues) is much larger than that of the other class (boundary residues), the Meta-Cost algorithm[4] can be used to improve prediction performance, in effect by increasing the penalty cost of missed boundary region predictions (false negatives). The "boundary" and "no boundary" signals were implemented as two neural network output units, from which the greater signal became the final prediction.

We used the SVM-Light[5] software implementation as our support vector machine classifier. A key decision in SVM classifier design is the kernel to use. For this work, we chose a polynomial kernel function. Since the outputs from the SVM classifier are positive or negative real numbers, the final prediction of a boundary was determined by an output signal greater than the estimated median.

For each protein chain predicted, a single output format was created. For each residue in the chain sequence, its amino acid was listed as were the boundary/non-boundary predictions for that position from both the ANN and the SVM. This gave a readily human-readable and interpretable format of the sequence and the boundary classification predictions. A human expert them made the final predictions based on his interpretation of the predictions of the two classification methods. The human prediction is the current output of the classification system.

The human expert consciously restricts his efforts to interpreting the machine predictions – *e.g.* filling in gaps in predicted boundaries, deleting short, unsupported boundary regions – rather than making a *de novo* prediction of the boundaries. The rationale for this is that we are currently implementing a mixture of experts model[6] to automate the combination of the ANN and SVM classifiers. The human expert provides the predictions until the mixture of experts model is finished, and will be a control against which to compare the mixture of experts model.

The human expert predictions we currently use are, of course, not absolutely blind, as the human may have previously seen PDB, CATH, *etc*. data for the predicted protein chain in the past, but a good faith effort is made to keep the predictions uncontaminated. For example, the human expert does not refer to any of the known structural data or other sources for the predicted protein chains.

We applied our system to a test set, containing 25 new sequences, and collected statistics for prediction accuracy. The artificial neural network by itself predicts a boundary region at a Q-observed accuracy of 37.50% and a correlation coefficient of 0.08 (Calculated as per Rost and Sander, 1993[7]). The support vector machine alone predicts a boundary region at a Q-observed accuracy of 76.75% and a correlation coefficient of 0.43. Finally, the application of a human expert yielded a Q-observed accuracy of 62.93% and a correlation coefficient of 0.41.

Our results show that protein structural domain boundaries can be predicted from amino acid sequence with respectable accuracy. Surprisingly, the best predictions were gotten using the SVM predictor alone. This beat both the ANN alone, and the human expert who was able to examine the predictions from both automated methods. The human expert, however, predicted more true positives than that of the SVM predictor, but the total accuracy was decreased by a larger amount of false positives. As we mentioned, we are in the process of replacing the human expert with an automated mixture of experts model to combine the results of the ANN and SVM classifiers.


1. Wernisch,L & Wodak,S.J. (2003). Identifying Structural Domains in Proteins. *Structural Bioinformatics*, P.E. Bourne, and H. Weissig (eds.). Wiley-Liss.
2. Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. & Thornton,J.M. (1997). CATH – a hierarchic classification of protein domain structures. *Structure* **5** (8), 1093-1109.
3. Rumelhart,D.E., Hinton,G.E. & Williams,R.J. (1986). Learning Internal Representations by Error Propagation. *Parallel Distributed Processing. Volume I: Foundations*, D.E. Rumelhart, J.L. McClelland and the PDP Research Groups (eds.). MIT Press.
4. Domingos,P. (1999) . MetaCost: A General Method for Making Classifiers Cost-Sensitive. Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining (pp. 155-164). San Diego, CA: ACM Press.
5. Joachims,J. (1999). Making Large-Scale SVM Learning Practical. *Advances in Kernel Methods – Support Vector Learning, B. Scholkopf, C. Burges &A. Smola (esd.), MIT-Pres.*
6. Duda,R.O., Hart,P.E. & Stark,D.G. (2001). Pattern Classification. Wiley-Interscience, New York.
7. Rost B. & Sander C. (1993). Prediction of Protein Secondary Structure at Better than 70% Accuracy. *J. Mol. Biol.* **232**(2), 584-599.

**UGA/IBM-PROSPECT** - 324 models for 64 3D / 26 FN targets

## Fold recognition using PROSPECT

J. Guo[1], W.J. Chung[1], K. Ellrott[1], R. Zhou[2], D. Gelonia[2], B.D. Silverman[2], A.K. Royyuru[2], A. Curioni[3], A. Logean[3], Y. Xu[1]

*[1] – Computational Systems Biology Laboratory, Department of Biochemistry and Molecular Biology, University of Georgia, Athens, GA, USA; [2] – Computational Biology Center, IBM Thomas J. Watson Research Center Yorktown Heights, NY, USA, [3] – Computational Biochemistry and Material Science, IBM Zurich Research Lab, 8003 Rueschlikon, Switzerland*

xyn@bmb.uga.edu

We have made predictions for all 76 valid targets using our protein structure prediction pipeline PROSPECT-PSPP[1]. The core of the pipeline is our newly improved fold recognition program, PROSPECT-III (manuscript under preparation), which solves the sequence-structure alignment problem using an integer programming method. Some of the top models from PROSPECT prediction were then screened using several different scoring functions and refined with the replica exchange molecular dynamics method (REMD).

PROSPECT employs both sequential and structural information for fold recognition and threading alignment. As in our previous version of PROSPECT[2], the evolutionary information is used not only in profile-profile sequence alignment score, but also in calculating the singleton and pair-wise energies, which greatly improves the performance on both fold recognition and alignment accuracy. Here, we employed an integer programming algorithm for finding an optimal threading alignment between a target sequence and structural templates measured by our energy functions. The advantage of using integer programming is that it can rigorously treat pair-wise and multi-body contact energy and allow variable gaps, and do so in a fairly efficient fashion in terms of actual computing time. A z-score is calculated for each optimal alignment through randomly shuffling the target sequence. The initial threading was done using a representative list from PISCES[3].

A typical prediction for each target starts with running the automatics prediction pipeline. If the prediction reliability is high and the quality of the model is good, as in most homology modeling cases, structural models were submitted with no or little human intervention. For other targets, additional information will be considered for template selection and alignment adjustments.

For some of the targets, the initial models were screened and refined if needed. Four different scoring functions were used in the screening process: (1) OPLSAA/PB energy, which is the minimized total energy of the protein using the OPLSAA force field and the Poisson-Boltzmann continuum solvent model; (2) Hydrophobic Score, which is defined as the surface area under the normalized second-order hydrophobic moment profile using an ellipsoidal description of protein shape[4]; (3) Correlation Score, the correlation coefficient between the distance of a residue from the center of the protein and its hydrophobicity; (4) mScore or mega Score, a combination of three scoring functions based on different grounds: a statistics based pairwise Cα-Cα distance dependant potential of mean force, a physics based non-bonded interaction energy of the GROMOS force field, and a phenomenological based Hydrophobic Score described above. Selected models were further refined with REMD method, which couples molecular dynamics trajectories with a temperature exchange Monte Carlo process for efficient sampling of the conformational space. In this method, replicas (total 12 in our implementation) are run in parallel at a sequence of temperatures ranging from the desired temperature to a high temperature at which the replica can easily surmount the energy barriers. From time to time the configurations of neighboring replicas are exchanged based on a Metropolis criterion. Because the high temperature replica can traverse high energy barriers, this provides a mechanism for the low temperature replicas to overcome the quasi-ergodicity they would otherwise encounter with a single temperature replica. The sampled conformations are then clustered and ranked based on the minimized total energy. The force field used in sampling is again the OPLSAA force field with the Poisson-Boltzmann continuum solvent model.

1. Guo,J-T. Ellrott,K. Chung,W.J. Xu,D. Passovets,S. Xu,Y. (2004). PROSPECT-PSPP: An Automatic Computational Pipeline for Protein Structure Prediction. *Nucleic Acids Res.* **32**, W522-525.
2. Kim,D. Xu,D. Guo,J-T. Ellrott,K. Xu,Y. (2003). PROSPECT II: Protein Structure Prediction Program for the Genome-scale Application. *Protein Eng.* **16**, 641-650.
3. Wang,G. Dunbrack,R.L. Jr. (2003). PISCES: a protein sequence culling server. *Bioinformatics* **19**, 1589-1591.
4. Zhou,R. Silverman,B.D. Royyuru,A.K. Atham,P. (2004). Spatial Profiling of Protein Hydrophobicity: Native vs. Decoy. *Proteins* **52**, 561-572.

**VENCLOVAS** - 25 models for 25 3D targets

## Comparative modeling by the consensus of sequence-structure mapping and structure assessment

Č. Venclovas and M. Margelevičius
*Institute of Biotechnology, Graičiūno 8, Vilnius, Lithuania*
venclovas@ibt.lt

In CASP6 we focused on those target proteins, for which evolutionary related proteins having known structure could be detected independently of the level of sequence similarity. In other words, our method could be classified as a template-based modeling.

Structural templates

PDB templates were identified by running either BLAST or PSI-BLAST[1] searches against the PDB sequence database or non-redundant NCBI sequence database respectively. If no significant matches to PDB entries were detected then consensus results reported by the GeneSilico fold recognition meta-server (http://genesilico.pl/meta)[2] were consulted to identify potential structural templates. If available, multiple templates usually were used to generate three-dimensional (3D) models. The selection of the structural templates was done attempting to represent observed conformational variations within the protein family/superfamily in an unbiased way.

Sequence-structure alignments

For high homology targets, where structural template(s) were among closely related sequences, alignments were derived directly from BLAST or PSI-BLAST results with some manual adjustments around insertions/deletions. For distant homology targets, two methods were used to generate and preliminary assess the alignment confidence in a region-specific manner. In the first method, results of an initial PSI-BLAST search were used in our intermediate sequence search procedure (PSI-BLAST-ISS)[3]. In this procedure, a set of sequences that bridge sequence space between target sequence and template(s) were used to initiate additional PSI-BLAST searches against the non-redundant sequence database. Target-template sequence alignments were then extracted from search results and their consistency was analyzed. For regions where one dominant alignment variant was produced, the alignment was considered reliable, while the regions where the consistency of target-template alignment was lacking were deemed unreliable. In the second method, publicly available 3D models for a particular target that were submitted to CASP6 by automatic servers were each superimposed with one of the templates using DaliLite[4]. Next, the structure-based multiple sequence alignment between the template and model sequences was constructed from obtained pairwise superpositions. The region-specific alignment reliability was then assessed as in the first method. Results by both methods were contrasted and consensus regions were considered to be reliably aligned. For the remaining regions alternative alignment variants were evaluated at the level of 3D models. Models based on these alternative alignments were assessed by several methods including ProsaII profiles and Z-scores[5], Verify3D profiles[6] and visual inspection. One of the main numerical indicators used to monitor the model quality upon evaluation of alternative alignments and exact placement of insertions/deletions was ProsaII Z-score, which was targeted to exceed the value for the best server-generated model.

Generating 3D structures

From given sequence-structure alignments models were generated automatically with MODELLER[7]. In most cases side chains were rebuilt using SCWRL[8]. No energy minimization procedures were used.

1. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-402.
2. Kurowski,M.A. & Bujnicki,J.M. (2003). GeneSilico protein structure prediction meta-server. *Nucleic Acids Res* **31**, 3305-7.
3. Venclovas,Č. (2001). Comparative modeling of CASP4 target proteins: combining results of sequence search with three-dimensional structure assessment. *Proteins* **Suppl 5**, 47-54.
4. Holm,L. & Sander,C. (1996). Mapping the protein universe. *Science* **273**, 595-603.
5. Sippl,M.J. (1993). Recognition of errors in three-dimensional structures of proteins. *Proteins* **17**, 355-62.
6. Luthy,R., Bowie,J.U. & Eisenberg,D. (1992). Assessment of protein models with three-dimensional profiles. *Nature* **356**, 83-5.
7. Šali,A. & Blundell,T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **234**, 779-815.
8. Canutescu,A.A., Shelenkov,A.A. & Dunbrack,R.L., Jr. (2003). A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci* **12**, 2001-14.

## **Wolynes_Schulten** - 81 models for 21 3D

### *Ab initio* structure prediction with associative memory Hamiltonians

M.C. Prentiss[1], C. Zong[1], G. Papioan[2], Z. Luthey-Schulten[3] and P.G. Wolynes[1]

*1– University of California, San Diego, 2– University of North Carolina, Chapel Hill, 3–University of Illinois, Urbana-Champaign*

pwolynes@chem.ucsd.edu

We initially selected sequences for ab initio prediction if there was no obvious scaffold found by automated comparative modeling servers. For the selected sequences, we used an Associative Memory Hamiltonian (AMH), with optimized parameters. The optimization procedure is used to pursue an energy landscape that discriminates the native state, while avoiding kinetic traps. The AMH energy function most often used in the submitted prediction included a nonpairwise additive potential based on solvent mediated interactions[1]. Different parameters have been optimized for proteins with all alpha and those with mixed all alpha-beta secondary structure units[2,3]. The alpha-beta energy function includes a sequence specific hydrogen bond term as well as a term that mimics the liquid crystal phase ordering of the beta strands[4]. We averaged the AMH potential over multiple sequence homologues when they were available. In most cases, information from secondary structure prediction was used to bias independent secondary structure units to their predicted structures. Molecular dynamics simulated annealing was used to select low energy candidate structures. Also constant temperature runs near the predicted folding temperature were used to generate candidate structures. Subsequently, a smaller subset of structures was selected for submission by evaluating the size of the hydrophobic core and the hydrophilic surface area. Further selection criteria included visual inspection, agreement with the preliminary secondary structure prediction and low energies predicted from a second optimized contact energy function.

1. Papoian,G.A. et al. (2004) Water in Protein Structure Prediction. *Proc. Nat. Acad. Sci. U.S.A.* **101**, 3352-3357.
2. Eastwood,M.P. et al. (2002) Statistical Mechanical Refinement of Protein Structure Prediction Schemes: Cumulant Expansion Approach. *J. Chem. Phys.* **117**, 4602-4615.
3. Hardin,C. et al. (2002) Associative Memory Hamiltonians for Structure Prediction Without Homology:Alpha-Beta Proteins. *Proc. Nat. Acad. Sci. U.S.A.* **100**, 1679-1684.
4. Hardin,C. et al. (2000) Associative Memory Hamiltonians for Structure Prediction Without Homology:Alpha-Helical Proteins. *Proc. Nat. Acad. Sci. U.S.A.* **97**, 14235-14240.

## **Wymore** - 32 models for 19 3D targets

### Comparative modeling using alternative alignments, statistical potentials and replica exchange simulations

Adam Marko, Stuart Pomerantz, Troy Wymore

*Biomedical Initiative Group, Pittsburgh Supercomputing Center, Pittsburgh, PA*

wymore@psc.edu

We have developed a protein structure prediction pipeline that is currently applicable for comparative modeling targets. This pipeline consisted of 1) generating hundreds of alternative alignments between target and template 2) using these alignments to generate structures 3) scoring these structures with a statistical potential and 4) visually examining lowest energy structures in an effort to pick the one closest to native. Programs were written in Perl to enable the flow between modeling programs. For variable regions in some targets we carried out a multi-scale modeling strategy combining lattice-based representations for sampling with all-atom models for ranking.

For all CASP6 targets, we first performed a BLAST[1] search through the non-redundant database. The sequences with significant E-values were collated and sequence profiles were constructed with the MEME[2] program. The MEME profiles were then used with the MAST[3] program to search for both additional sequences and structural templates. In a few instances we used 3D-PSSM[4] to help identify templates. If we were satisfied with the list of related sequences and structural template(s), we performed a multiple sequence alignment with the T-coffee[5] program to make an initial determination on the level of difficulty in modeling the structure.

Based on the T-coffee alignment and time constraints, we constructed 100-500 alternative alignments between template and target using the program probA[6]. This program uses a probabilistic backtracking procedure that generates ensembles of suboptimal alignments with correct statistical weights. This ensemble of alignments and the one from the T-coffee program were used to build structures using MODELLER version 6.2[7]. The structures were then scored using the sum of CA-CA, CB-CB and surface statistical potentials in ProsaII[8]. Typically, up to 20 of the lower energy models were visually examined with a graphics program. For some target predictions that were not subsequently refined, we would minimize several structures identified as favorable through ProsaII with an all-atom molecular mechanical (MM) distance-dependant dielectric potential and then score them with an all-atom MM-Generalized Born potential using the MMTSB[9] toolset and submit the lowest energy structure. Almost all predicted structures were minimized with restraints using the MMTSB toolset.

For six highly variable regions ranging in size from 5-16 residues, we performed lattice-based replica exchange simulations using MONSSTER[10] through the MMTSB toolset. The lowest temperature structures from the final rounds of simulation (typically the last 100-1000 structures) were rebuilt to complete all-atom models and clustered according to distance RMSD. The clusters were minimized and ranked with the same potentials as described above. Finally we would generally choose the lowest energy structure from the cluster exhibiting the lowest average energy or from a highly populated low energy cluster. All modeling tasks were greatly facilitated by use of the MMTSB toolset. Our goals were to demonstrate improvement in our comparative models over those constructed from a T-coffee alignment and assess our sampling and scoring procedures both in sequence and 3D space.


1. Altschul,S.F., Gish,W., Miller,W., Meyers,E.W., Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410.
2. Bailey,T.L., Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. 2nd Int. Conf. Intelligent Sys. Mol. Biol.* AAAI Press, 28-34.
3. Bailey,T.L., Gribskov,M. (1998) Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* **14**, 48-54.
4. Kelley,L.A., MacCallum,R.M., Sternberg,M.J.E. (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.* **299**, 499-520.
5. Notredame,C., Higgins,D., Heringa,J. (2000) T-Coffee: A novel method for multiple sequence alignments. *J. Mol. Biol.*, **302**, 205-217.
6. Muckstein,U., Hofacker,I.L., Stadler,P.F. (2002) Stochastic pairwise alignments. *Bioinformatics*, **18**, S153-S160.
7. Sali,A., Blundell,T.L. (1993) Comparative Protein Modeling by Satisfaction of Spatial Restraints. *J. Mol. Biol*., **234**,779-815.
8. Sippl,M.J. (1993) Recognition of Errors in Three-Dimensional Structures of Proteins. *PROTEINS: Struct. Func. Gen.* **17**,355-362.
9. Feig,M., Karanicolas,J., Brooks III,C.L.B. (2004) MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology. *J. Mol. Graph. Model.* **22**, 377-395.
10. Skolnick,J., Kolinski,A., Ortiz,A.R. (1997) MONSSTER: a method for folding globular proteins with a small number of distance restraints. *J. Mol. Biol.* **265**, 217-241.


## **YASARA** - 28 models for 9 3D targets

### WHAT IF YASARA folds a protein?


E. Krieger, S.B. Nabuurs, C.A.E.M. Spronk and G. Vriend
*CMBI, Center for Molecular and Biomolecular Informatics,*
*Radboud University Nijmegen, the Netherlands*
Elmar.Krieger@cmbi.ru.nl, www.YASARA.org


The *homology modeling module of the YASARA/WHAT IF Twinset* integrates functions provided by both programs and a variety of fold recognition and secondary structure prediction servers into a fully automatic method for protein structure prediction.

*Initial alignments were collected from the 3D-Jury system* on the CAFASP website[1]. A consensus secondary structure prediction was obtained from PSIPRED[2] and SAM-T02[3]. Alignments were pooled and sent through secondary- and tertiary structure-based correction filters. Loops and structured N- and C-termini were added with YASARA's loop modeler, side-chains were completed by WHAT IF[4] and SCWRL[5]. These models were submitted with the 'UNREFINED' keyword.

*In the refinement stage*, the conformational space available to the models was sampled with CONCOORD[6], then *hundreds of all-atom molecular dynamics simulations* in aqueous solution (Particle Mesh Ewald electrostatics[7]) were run with YASARA to 'home in' further to the target. This was done with the new *YASARA force field*, a third-generation self-parameterizing energy function[8] obtained in crystal space from the YAMBER force field[9]. Models were ranked based on the WHAT IF/YASARA ColonyMorphScore, a scoring function that combines various checks done by WHAT IF[10] with the energy assigned by the YASARA force field.

Due to the huge computational requirements, the entire procedure was run in parallel using the Models@Home distributed computing system[11]. Thanks to everyone working here at the CMBI in Nijmegen, Netherlands, for choosing the Models@Home screensaver.

More information is available at *www.yasara.org* and *www.cmbi.ru.nl/whatif*


1. von Grotthuss,M., Pas,J., Wyrwicz,L., Ginalski,K. & Rychlewski,L. (2003) Application of 3D-Jury, GRDB, and Verify3D in fold recognition. *Proteins* **53, Suppl. 6**, 418-423.
2. McGuffin,L. J., Bryson,K. & Jones,D.T. (2000) The PSIPRED protein structure prediction server. *Bioinformatics* **16**, 404-405.
3. Karplus,K. et al. (1999) Predicting protein structure using only sequence information. *Proteins* **37(S3)**, 121-125.
4. Chinea,G., Padron,G., Hooft,R.W.W., Sander,C. & Vriend,G. (1995) The use of position specific rotamers in model building by homology. *Proteins* **23**, 415-421.
5. Canutescu,A.A., Shelenkov,A.A. & Dunbrack,R.L.J. (2003) A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci.* **12**, 2001-2014.
6. de Groot,B.L. et al. (1997) Prediction of protein conformational freedom from distance constraints. *Proteins* **29**, 240-251.
7. Essman,U. et al. (1995) A smooth particle mesh Ewald method. *J.Chem.Phys.* **103**, 8577-8593.
8. Krieger,E., Koraimann,G. & Vriend,G. (2002) Increasing the precision of comparative models with YASARA NOVA - a self-parameterizing force field. *Proteins* **47**, 393-402.
9. Krieger,E., Darden,T., Nabuurs,S.B., Finkelstein,A. & Vriend,G. (2004) Making optimal use of empirical energy functions: force field parametrization in crystal space. *Proteins* **in press**.
10. Hooft,R.W.W., Vriend,G., Sander,C. & Abola,E.E. (1996) Errors in protein structures. *Nature* **381**, 272-272.
11. Krieger,E. & Vriend,G. (2002) Models@Home: distributed computing in bioinformatics using a screensaver based approach. *Bioinformatics* **18**, 315-318.


## **Zhou-SP$^3$** (serv) - 320 models for 64 3D targets

### Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments in CASP6


Hongyi Zhou and Yaoqi Zhou
*HHMI Center for Single Molecule Biophysics*
*Department of Physiology & Biophysics, State University of New York*
*Buffalo, NY 14214*
hzhou2@buffalo.edu, yqzhou@buffalo.edu


Recognizing the structural similarity of proteins without significant sequence identity (fold recognition) has proven to be a challenging task. One way to detect structural similarity is to identify remote sequence homology via sequence comparison. Advances have been made from the pairwise to multiple sequence comparison, from sequence-to-sequence, sequence-to-profile to profile-to-profile comparison. Another way to detect structural similarity is to take full advantage of known protein structures. For example, the sequence-to-structure threading assesses the compatibility of a sequence with each known structure by a pairwise score function or single-body structural profile. In recent work, attempts were made to optimally combine the sequence and structure information for a more accurate/sensitive fold recognition. Most focused on combining sequence information with threading techniques.

One intuitive approach to incorporate structural information is to use structural alignment. Application of structural alignment to fold recognition has been mostly limited to the derivation of substitution matrices. The direct incorporation of sequence profiles generated from structural alignment, however, does not appear to be useful for remote homology detection. For example, Gough et al.[1] found that hidden Markov models (HMM) generated from structural alignment yielded poorer results than HMMs generated independently. Tang et al.[2] showed that the combination of sequence profiles derived from structural alignments for protein-core regions with the sequence profiles from sequence alignment and secondary structural profiles does not further improve fold recognition sensitivity by profile-profile alignment. This

highlights the difficulty of harnessing structural information in a combined approach for optimal fold-recognition alignment[3,4]. In fact, recently completed LiveBench 8 and 9 tests suggested that the top performers of the fold-recognition servers of single methods are sequence-based profile-profile alignment methods such as BasD/mBas/BasP, SFST/STMP, FFAS03, and ORFeus/ORFeus2. Not a single method using structural information was made to the top four[5].

We are developing a novel, combined approach based on sequence profiles generated from structural alignment. In this approach, fragments rather than whole proteins are used for structural alignment. The use of fragments has following advantages over the use of whole protein for structural alignment. First, there is a sufficient coverage for all possible structures of short fragments in the existing structures in protein data bank. The large number of fragments contained in protein data bank leads to a statistically significant sequence profile. In contrast, sequence profiles generated from structural alignment of whole proteins require that all proteins have a sufficient number of structurally similar proteins with low sequence identity – a condition that is difficult to meet. Second, the use of fragments allows producing a reliable sequence profile for all regions of a protein. In structural alignment of whole proteins, however, many regions (loop regions, in particular) are not aligned. Third, unlike structural alignment of proteins, structural alignment of fragments is more likely to have a unique solution because their structural topologies are relatively simple.

Another unique feature of the new approach is that alignment of two fragments is not only characterized by their structural difference (rmsd) but also by their positions from solvent (residue depth). This partially remedies the loss of the information on the environment surrounding the fragments.

The sequence profile (SP) derived from depth-dependent structure alignment of fragments allows a simple integration with evolution-derived sequence profile (SP) and secondary-structural profile (SP) for an optimized fold-recognition alignment by efficient local-local dynamic programming, secondary-structure-dependent gap penalty, and a sophisticate empirical ranking method. The resulting method (called $SP^3$) is found to make a statistically significant improvement in both the sensitivity of fold recognition and the accuracy of alignment compared to the method based on evolution-derived sequence profiles alone (SP) and the method based on evolution-derived sequence profile and secondary structure profile ($SP^2$). $SP^3$ was tested in SALIGN benchmark for alignment accuracy[6] and Lindahl[4], PROSPECTOR 3.0[7], and LiveBench 8.0 benchmarks[5] for remote-homology detection and model accuracy. $SP^3$ is found to be the most sensitive and accurate single-method server in all benchmarks tested where other methods are available for comparison (although its results are statistically indistinguishable from the next best in some cases and the comparison is subjected to the limitation of time-dependent sequence and/or structural library used by different methods.). $SP^3$ participates CASP6 as a server located at http://theory.med.buffalo.edu. The new approach proposed here hopefully will stimulate more new ideas in attacking the challenging fold recognition problem.


1. Gough,J., Karplus,K., Hughey,R., Chothia,C. (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.* **313,** 903-919.
2. Tang,C.L., Xie,L., Koh,I.Y., Posy,S., Alexov,E., Honig,B. (2003) On the role of structural information in remote homology detection and sequence alignment: new methods using hybrid sequence profiles. *J. Mol. Biol*. **334**, 1043–1062.
3. Panchenko,A.R, Marchler-Bauer,A.,Bryant,S.H., (2000) Combination of threading potentials and sequence profiles improves fold recognition. *J. Mol. Biol*. **296,** 1319–1331.
4. Lindahl,E., Elofsson,A., (2000) Identification of related proteins on family, superfamily and fold level. *J. Mol. Biol.* **295**, 613–625.
5. http://BioInfo.PL; Bujnicki,J.M., Elofsson,A., Fischer,D., Rychlewski,L. (2001) Livebench-1: Large-scale automated evaluation of protein structure prediction servers. *Protein Sci.* **10**, 352–361.
6. Marti-Renom,M.A, Madhusudhan,M., Sali.A. (2004) Alignment of protein sequences by their profiles. *Protein Sci.* **13**, 1071–1087
7. Skolnick,J., Kihara,D., Zhang,Y. (2004) Development and large scale benchmark testing of the PROSPECTOR 3.0 threading algorithm. *Proteins* **55**, 502-518.


## **Zhou-SPARKS2** (serv) - 320 models for 64 3D targets

### Application of SPARKS 2.0 fold recognition server in CASP6


Hongyi Zhou and Yaoqi Zhou
*HHMI Center for Single Molecule Biophysics*
*Department of Physiology & Biophysics, State University of New York*
*Buffalo, NY 14214*
hzhou2@buffalo.edu,yqzhou@buffalo.edu


SPARKS is a method that combines Sequence, secondary structure Profiles with A single-body Residue-level Knowledge-based Score for fold recognition.

While there exist many fold recognition methods that integrate sequence information with threading techniques, SPARKS uses an elaborate knowledge-based score that contains a torsion-angle term for backbone interaction and a combined buried surface and contact-energy term for residue-residue and residue-solvent interactions. Most other methods used a much simpler single-body or profile energy score that takes into account of solvent exposure or contact score only. The use of a single-body energy score allows the use of efficient dynamic programming for optimal fold-recognition alignment. This is in contrast to the pairwise score function which would require extensive computing time and/or additional approximations (such as frozen approximation) for optimal alignment. SPARKS was tested in ProSup, Lindahl benchmark and LiveBench[1].

SPARKS 2.0 improves over SPARKS in following areas. First, a local-local dynamic programming rather than a global-local method is used. This improves the alignment between two sequences with very different sequence lengths. Second, a gap penalty that depends on the secondary structure is introduced. This allows a more accurate sequence alignment. Third, an empirical method for ranking templates is introduced. This method uses 1) the difference between an alignment score of the query sequence and a template sequence and that of the reversed query sequence and the template, 2) structural similarity score between top-ranked models, and 3) scores normalized by "true" alignment length (excluding ending gaps) and full alignment lengths (including all gaps). SPARKS 2.0 was tested on SALIGN benchmark for alignment accuracy[2] and Lindahl[3], PROSPECTOR 3.0[4], and LiveBench 8.0 benchmarks[5] for remote-homology detection and model accuracy. The accuracy and sensitivity of SPARKS 2.0 (Table I) were found to be slightly worse than SP[3] (a different fold recognition method that harnesses structural information without the need of threading, see a separate abstract by Zhou and Zhou). However, it is one of the best methods that combine sequence profiles with threading for fold recognition. Thus, its result will be useful for consensus methods.

Table I. The results of various benchmark testing.

| Method | SALIGN[2] | Lindahal[3] | | PROSPECTOR 3.0[4] | | LiveBench 8 | |
|---|---|---|---|---|---|---|---|
| | AlignAccu. | Accu.[a] | Sens.[b] | Accu.[a] | Sens.[b] | Accu.[a] | Sens.[b] |
| Other[c] | 56.4%[d] | | | 520.1[e] | 925[e] | 41.91[f] | 112[f] |
| SPARKS | 53.1% | 325.9 | 611 | 529.0 | 979 | 38.33 | 99 |
| SPARKS2 | 54.9% | 349.2 | 655 | 591.0 | 1041 | 40.7 | 119 |
| SP[3] | 56.6% | 349.2 | 665 | 601.9 | 1066 | 42.2 | 120 |

[a] Accuracy by total MaxSub score of first ranked models.

[b] Sensitivity by number of models with MaxSub score >0.01.

[c]The known best performance in the corresponding benchmark by other methods.

[d]SALIGN is trained and tested by CE alignment.[2] SPARKS and SP3 are not trained by CE alignment. Thus, matching the performance of SALIGN by SPARKS2 and SP3 is remarkable.

[e]PROSPECTOR 3.0

[f]The best single method server (BasD). Other consensus methods such as SHOTGUN have higher accuracy and sensitivity.


1. Zhou,H., Zhou,Y. (2004) Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. *Proteins* **55**, 1005-1013.
2. Marti-Renom,M.A., Madhusudhan,M., Sali.A. (2004) Alignment of protein sequences by their profiles. *Protein Sci.* **13**, 1071–1087.
3. Lindahl,E., Elofsson,A. (2000) Identification of related proteins on family, superfamily and fold level. *J. Mol. Biol.* **295**, 613–625.
4. Skolnick,J., Kihara,D., Zhang,Y. (2004) Development and large scale benchmark testing of the PROSPECTOR 3.0 threading algorithm. *Proteins* **55**, 502-518.
5. http://BioInfo.PL; Bujnicki,J.M., Elofsson,A., Fischer,D., Rychlewski,L. (2001) Livebench-1: Large-scale automated evaluation of protein structure prediction servers. *Protein Sci.* **10**, 352–361.

## **Accelrys** - 27 models for 16 3D / 1 FN targets

### ChiRotor and Looper for side-chain and loop optimization

D. Singh, Taisung Lee, V. Spassov, L. Yan and D. Haley-Vicente
*Accelrys Inc., 9685 Scranton Rd., San Diego, CA 92121*
dhv@accelrys.com

CASP6 target homology models were predicted using a suite of tools available in Discovery Studio® (DS) Modeling and Insight II® modeling and simulations packages (Accelrys, Inc)[1,2]. Several models have been further optimized by two new methods developed at Accelrys, ChiRotor and Looper[3,4], for side-chain and loop optimization, respectively. ChiRotor is a fast conformational search algorithm that combines rotamer searches with an energy evaluation to calculate optimal side-chain conformations for all or part of a protein with an average RMSD of ~1Å for the core residues. Looper is an fast algorithm that performs a hierarchical search of low energy loop structures to provide a ranked list of loop fragments with a high level of accuracy.

1. Discovery Studio Modeling (http://www.accelrys.com/dstudio/ds_modeling/) Accelrys Inc.
2. Insight II (http://www.accelrys.com/insight/) Accelrys Inc.
3. Spassov,V.Z., Yan,L. (2004) ChiRotor: A side-chain prediction algorithm based on side-chain backbone interactions. *To be submitted.*
4. Spassov V.Z., Yan,L. (2004) Looper: A CHARMm based algorithm for loop prediction using hierarchical structural optimisation. *In preparation.*

## **BAKER** - 433 models for 64 3D / 63 RR / 58 FN

### Novel approaches to protein structure prediction at CASP6

P. Bradley, G. Cheng, D. Chivian, D. Kim, L. Malmstrom, J. Meiler, K. Misura, Bin Qian, J. Schonbrun, A. Zanghellini, D. Baker*
*University of Washington*
dabaker@u.washington.edu

See methods section

### High resolution refinement can be successful in low dimensional search spaces

Bin Qian, Ora Furman, Chu Wang and David Baker
*University of Washington*
dabaker@u.washington.edu

Accurate high-resolution refinement of protein structure models is a formidable challenge because of the delicate balance of forces in the native state, the difficulty in sampling the very large number of alternative tightly packed conformations, and the inaccuracies incurrent force fields. Indeed, energy-based refinement of comparative models generally leads to degradation rather than improvement in model quality, and hence, most current comparative modeling procedures omit physically based refinement. However, despite their inaccuracies, current force fields do contain information that is orthogonal to the evolutionary information on which comparative models are based, and hence, refinement might be able to improve comparative models if the space that is sampled is restricted sufficiently so that false attractors are avoided. We have found that full atom refinement canimprove model accuracy both in high resolution protein-protein docking calculations where the space searched consists of the sidechain torsion angles and the rigid body degrees of freedom, and in comparative model refinement where the search space is defined by side chain torsion angles and variation of the backbone along evolutionarily favored sampling directions given by the principal components of the variation of backbone structures within a homologous family.

### Membrane Protein Structure Prediction with ROSETTA

Jack Schonbrun, Vladimir Yarovoy, David Baker
*University of Washington*
dabaker@u.washington.edu

Membrane proteins are the among the most important targets for structure prediction. Though they number less than 1% of the structures in the Protein Databank, they are estimated to account for 20-30% of all open reading frames. In an attempt to model these proteins, we have derived a new statistical potential and implemented it in ROSETTA. We calculated environment dependent amino acid propensities based on observed frequencies in a set of solved membrane protein structures. The environment of a residue is now a function of two parameters: depth within the membrane, and number of

neighbors. The score and search are also modified to bias the results toward helical bundles. By taking advantage of the mostly local topologies of known structures, we have had some success is producing low-resolution native like models.

## **BAKER-ROBETTA** (serv) - 320 models for 64 3D targets
## **BAKER-ROBETTA_04** - 320 models for 64 3D targets

### The Robetta and Robetta_04 protocols

Dylan Chivian[1], David E. Kim[1], Lars Malmstrom[1],
Jack Schonbrun[1], Carol A. Rohl[2] & David Baker[1,*]
*1- University of Washington, Seattle, WA*
*2- University of California, Santa Cruz, CA*
dabaker@u.washington.edu

See methods section

## **BAKER-ROBETTA-GINZU** (serv) - 64 models for 64 DP targets

### The Ginzu homologue identification and domain parsing protocol

Dylan Chivian, David E. Kim, Lars Malmstrom, & David Baker[*]
*University of Washington, Seattle, WA*
dabaker@u.washington.edu

See methods section

## **BAKER-ROSETTADOM** (serv) - 64 models for 64 DP targets

### The RosettaDOM domain parsing protocol

D.E. Kim, D. Chivian, L. Malmstrom and D. Baker
*University of Washington*
dabaker@u.washington.edu

See methods section

## **Casplta** - 348 models for 64 3D / 63 DP / 63 DR / 64 FN

### The Victor/FRST function for model quality estimation

S.C.E. Tosatto[1]
*[1] – Dept. of Biology and CRIBI Biotech Centre, University of Padova*
silvio@cribi.unipd.it

The Victor/FRST (Function of Rapdf, Solvation and Torsion potentials) function is a statistical scoring function used to estimate the quality of a protein structure. It is implemented as the weighted linear combination of three different components covering the major aspects of structure quality estimation.

The first component is an implementation of the RAPDF[1] statistical pairwise potential. This potential of mean force discriminates between residue specific non-bonded interactions at the atomic level, e.g. the $C_\alpha$ of an Isoleucine is a different type from the $C_\alpha$ of a Glycine. It is used with published parameters. A simple solvation potential is derived in analogy to the one described for GenTHREADER[3]. The relative solvent accessibility is estimated as the number of other $C_\beta$ atoms within a sphere of radius 10 A centered on the residue's $C_\beta$ atom. The reference state for this distribution is generated from the TOP500 database[2]. This database of high resolution crystal structures is used to estimate the relative probability of encountering a number $i$ ($i$ = 0,…,40) of $C_\beta$ atoms surrounding each of the 20 amino acids. The energy for a given structure is calculated with the standard log scale for mean force potentials. A similar scheme was also used to parameterize the torsion angle potential. All ($\phi$,$\psi$) angle combinations, discretized in 10x10 degree bins, present in the TOP500 database[2] are used to estimate the reference state for each of the 20 amino acids. The same log scale formula is applied to derive an energy for a given structure.

Since the three components have different orders of magnitude and cannot be related directly to the same scale, weighting factors are used before summing the partial energies. These factors were optimized on the CASP-4 decoy set[4] optimizing the linear correlation between total energy and GDT_TS score[5] as target function. The final scoring fuction was submitted to CAFASP-4 in the MQAP (Model Quality Assessment Program) category.


1. Samudrala,R., & Moult,J. (1998) An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J. Mol. Biol.* **275**, 895-916.
2. Lovell,S.C., Davis,I.W., Arendall,W.B.r., de Bakker,P.I., Word,J.M., Prisant,M.G., Richardson,J.S., & Richardson,D.C. (2003) Structure validation by Calpha geometry: phi,psi and Cbeta deviation. *Proteins* **50**, 437-450.
3. Jones,D.T. (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* **287**, 797-815.
4. URL:*http://predictioncenter.llnl.gov/download_area/CASP4/MODELS_SUBMITTED/*
5. Zemla,A. (2003) LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res*. **31**, 3370-3374.


## **CBRC-3D** - 319 models for 64 3D / 22 FN targets

### Comparative modeling and fold recognition using FORTE series


K. Tomii, T. Hirokawa, and C. Motono
*Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, 2-43 Aomi, Koto-ku, Tokyo, Japan*
k-tomii@aist.go.jp


See methods section

## **CHIMERA** - 65 models for 64 3D targets

### A versatile web user interface system for highly accurate protein structure prediction: SKE (Sophia-kai-Ergon) CHIMERA


M. Takeda-Shitaka, G. Terashi, D. Takaya, K. Kanou, M. Iwadate and H. Umeyama
*Kitasato University*
shitakam@pharm.kitasato-u.ac.jp


See methods section

## **FAMD** (serv) - 320 models for 64 3D targets

### Full automatic homology-modeling servers including wisdom and practice: SKE(Sophia Kai Ergon) FAMD


K. Kanou[1], M. Iwadate[1], G. Terashi[1], D. Takaya[1], M. Takeda-Shitaka[1] and H. Umeyama[1]
*[1] - Department of Biomolecular Desig*
*School of Pharmaceutical Sciences, Kitasato University*
kanouk@pharm.kitasato-u.ac.jp


See methods section

## **FAMS** (serv) - 320 models for 64 3D targets

### Full automatic homology modeling server including the transformation of amino acid residues: SKE(Sophia Kai Ergon) FAMS

M. Iwadate[1], K. Kanou[1], G. Terashi[1], D. Takaya[1],
M. Takeda-Shitaka[1] and H. Umeyama[1]
[1] - *Department of Biomolecular Desig*
*School of Pharmaceutical Sciences, Kitasato University*
iwadatem@pharm.kitasato-u.ac.jp

See methods section

## **Hamilton-Huber-Torda** (serv) - 61 models for 61 RR targets

### Protein contact prediction using patterns of correlation

N.A. Hamilton[1,2], K. Burrage[1], M.A. Ragan[2], A.E. Torda[3]
and T. Huber[1]
[1]*– Advanced Computational Modelling Centre, The University of Queensland,*
[2]*– Institute for Molecular Bioscience, The University of Queensland,*
[3] *– Zentrum für Bioinformatik, Universität Hamburg*
n.hamilton@imb.uq.edu.au

See methods section

## **HOGUE-HOMTRAJ** (serv) - 105 models for 45 3D targets

### HomTraj: an automated structure rediction server with a performance-monitoring test suite

K.A. Snyder[1], H.J. Feldman[1], F. Wu[1] and C.W.V Hogue[1,2]
[1] *– The Blueprint Initiative, Mount Sinai Hospital, Toronto, Canada,*
[2] *– Department of Biochemistry, University of Toronto, Toronto, Canada*
chogue@blueprint.org

HomTraj is a powerful fully automated web-based Homology Modeling prediction service that will return up to five structure predictions from a given query protein sequence. First, NCBI BLAST[1] (expect value 1e-20) is used to identify homologous templates from the PDB. If this call fails, the Sequence Alignment and Modeling (SAM2k) algorithm[2] is used to identify more remotely homologous structure templates from the PDB. The algorithm uses a two-track Hidden Markov Model (HMM) to identify homology – one track for sequence and one for secondary structure. A PsiPred[3] secondary structure prediction is used as input for the secondary HMM track.

Next, using a modified version of our TRADES algorithm[4], the backbone alpha-carbon trajectory of the template was recorded, and a trajectory distribution built with the new sequence of the target. Each gapless stretch of alignment was replaced by a single fragment from the recorded trace. Where gaps occurred in the alignment, fragments were built to span the gaps. Gaps may be shifted a few residues left or right in order to minimize the energy of the loop spanning the gap. Roughly 50 structures were generated using the fragments obtained from the previous steps and our Foldtraj software, with bump checking slightly reduced. Using a modified version of a statistical residue-based potential[5], which we have termed "crease energy", the best structure generated from each template was chosen and submitted. Structures can be provided in either PDB or NCBI ASN.1 format.

In an effort to quantify the performance of HomTraj on a diverse group of query proteins, a web-based test suite was recently developed. Using the test suites' web interface a user may customize a program run, selecting the appropriate HomTraj version and query test set. In addition, results from all previous runs can be accessed for analysis.

Three query test sets, easy, medium and hard, were generated in order to analyze the performance of HomTraj where the degree of template sequence homology to the query varies from very high to very low. Test sets were constructed from a set of 75 domains representative of diverse fold categories from the ASTRAL SCOP 1.65 Genetic Domain Sequences database[6]. Domains were subdivided into 3 levels of difficulty according to the E-value of top PDB template alignments returned from BLAST hits. Queries with BLAST hit E-values less than or equal to 1e-20 were classified as easy, those with E-values less than 1e-3 and greater than 1e-20 were classified as medium and those with E-values greater than 1e-3 were classified as hard. To enable efficient searching for structure templates by HomTraj a static version of NCBI's non-redundant PDB database was generated by removing proteins with a high level of sequence homology to the test set queries.

During each test suite run, results from HomTraj's subprograms are stored in MySQL tables to enable fast and efficient access from the web-interface. Upon completion of a test set, RMSD scores for each query, as well as an average RMSD, is displayed to facilitate comparisons between different versions of HomTraj. In this way, optimizations to the server can be accurately assessed and quantified.


1. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
2. Karplus,K., Karchin,R., Barrett,C., Tu,S., Cline,M., Diekhans,M., Grate,L., Casper,J. & Hughey,R. (2001). What is the value added by human intervention in protein structure prediction? *Proteins* **Suppl. 5,** 86-91.
3. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol. **292***, 195-202.
4. Feldman,H.J. & Hogue, C.W.V. (2000). A Fast Method to Sample Real Protein Conformational Space. *Proteins* **39,** 112-131.
5. Bryant,S.H. & Lawrence,C.E. (1993). An Empirical Energy Function for Threading Protein Sequence through the Folding Motif. *Proteins* **16,** 92-112.
6. Chandonia,J.M, Hon G., Walker N.S., Lo Conte L., Koehl P., Levitt M., & Brenner S.E. (2004). The ASTRAL compendium in 2004. *Nucleic Acids Res.* **32,** D189-D192.


## **Huber-Torda** - 242 models for 63 3D / 60 RR targets

### Sequence to structure alignments with fragment compatibility terms and an optimized substitution matrix


T. Huber[1], T. Lai[2], E. Mittag[2], J.B. Procter[2], H. Stehr[2], S. Mühlenmeister[2], B. Otto,[2] A.E. Torda[2]

*[1] – Dept of Mathematics, University of Queensland, Brisbane, Australia,*
*[2] –ZBH, University of Hamburg, Bundesstr 43, D-20146, Hamburg, Germany*
torda@zbh.uni-hamburg.de


The methods used in the "wurst" server[1], combine the most mundane elements of protein threading with some more entertaining ideas from score functions and alignments.[2] The methods contain structure-based terms, but are free of Boltzmann-based or *z*-score derived methods. These are combined with a sequence based-term, but without standard substitution matrices.

The structure-based terms rely on a sequence to local-structure compatibility function. To parameterise the term, more than $10^5$ fragments of length 9 were described by a set of continuous descriptors (for structural and solvation properties) and discrete descriptors (for sequence). A classification across both kinds of descriptor reduced this to a set of less than 500 classes. This is unlike other fragment libraries in the literature in that two classes may be structurally similar, but differ in sequence patterns. Because the classification method[3] is based on Bayesian statistics, it directly provides a log-odds probabilities and is easy to convert to a scoring matrix.

The next component was a sequence-based term using an unusual substitution matrix. A classic simplex optimization method was used to adjust the 210 members of a substitution matrix using a cost function which measured the quality of alignments or more specifically, the quality of the models produced by alignments. For this parameterization, a calibration set of proteins was collected consisting of pairs of similar structures with low sequence identity. Each protein's sequence was aligned against that of its partner and the resulting model compared to the original (correct) structure. The better the quality of the model, the lower the cost function as summed over the calibration set.

Alignments were calculated using a standard dynamic programming method applied to a matrix which was a linear combination of the sequence- and structure-based terms, but with still more optimizations at the parameterization stage. The final parameters used for CASP resulted from a simultaneous optimization of the weighting of the two main terms, the gap penalties and the elements of substitution matrix. Furthermore, the optimization was done using sequence profiles rather than the sequences to be aligned

The net result is has some unusual properties. The substitution matrix is very different to a BLOSUM matrix in that it has more weight on diagonal terms, since it is optimized for profiles. The matrix used for CASP is even more unusual in that it is adjusted to work best in the field due to the structural terms. The final result is the machinery for producing very good sequence to structure alignments in the face of low sequence identity.

Preliminary results already show some strengths and weaknesses of the approach. The optimization procedure is very effective, but had an unexpected side effect. The parameters were very tightly tuned to the properties of the sequence profiles and produced poor alignments if a sequence did not have some number of close sequence homologues. The framework used to create the structural term is very well suited to sequence to structure alignments, but the structural descriptors we chose are probably still not ideal. Finally, a huge

weakness was in the ranking of the produced models and failing to reasonably account for effects of sequence, structure and model size. Several good models were produced, but only ranked in the top 10 to 20 guesses rather than first place.

Some of these weaknesses have been repaired (after CASP). Some leave us with something to do for the next year.


1. http://www.zbh.uni-hamburg/wurst
2. Torda,A.E.,Procter,J.B. & Huber,T. (2004) Wurst: A protein threading server with a structural scoring function, sequence profiles and optimised substitution matrices. *Nucl. Acids Res.* **32**, W532-W535.
3. Cheeseman, P. & Stutz, J., Bayesian classification (autoclass): Theory and results, in Advances in knowledge discovery and data mining, U. Fayyad, et al., Editors. 1995, The AAAI Press: Menlo Park. p. 61-83.


## **IUPred** - 57 models for 56 DR targets

### Prediction of protein disorder based on the estimation of pairwise interaction energy


Zsuzsanna Dosztányi, Veronika Csizmók, Péter Tompa and István Simon

*Institute of Enzymology, Biological Research Center, Hungarian Academy of Science, Budapest, Hungary*

zsuzsa@enzim.hu


See methods section

## **Jones-UCL** - 251 models for 63 3D / 64 DR / 26 FN

### Improving the Quality of Fold Recognition Models Using the nFOLD Method


L.J. McGuffin, J.S. Sodhi, K. Bryson and D.T. Jones

*-Bioinformatics Unit, Department of Computer Science, University College London, London WC1H 6BT*

l.mcguffin@cs.ucl.ac.uk


We have developed a new fold recognition method, nFOLD, that extends the new profile-profile version of mGenTHREADER[1,2], through the incorporation of a number of extra inputs into the underlying neural network.

Three additional inputs are fed into the neural network which include: the secondary structure element alignment (SSEA) score[2], a new functional site detection score (MetSite)[3] and a simple model quality checking algorithm, MODCHECK[4]. The nFOLD neural network is also trained directly on MaxSub[5] score which allows for a greater assignment of confidence in model quality. Although the SSEA score has been benchmarked previously as an extra neural network input to mGenTHREADER[2], this is the first time it has been included in a fully automated method within a blind assessment.

The functional site predictions were calculated using a set of classifiers based on the MetSite method[3], which was initially developed in order to predict the location of residues forming commonly occurring metal binding sites in low-resolution structural models. The top ranking MetSite predictions were extracted for the models generated from the mGenTHREADER profile-profile alignments. Analysis of the MetSite scores showed a significant improvement in distinguishing native and near native-like models from decoy hits.

The MODCHECK score was also used to directly assess the quality of the models from the profile-profile alignments. The MODCHECK program has been used previously for our CASP predictions[4], however this is the first time it has been implemented in a fully automated method.


1. Jones,D.T. (1999) GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* **287**, 797-815.
2. McGuffin,L.J. & Jones,D.T. (2003) Improvement of the GenTHREADER method for genomic fold recognition. *Bioinformatics*. **19**, 874-881.

3. Sodhi,J.S., Bryson,K., McGuffin,L.J., Ward,J.J., Wernisch,L. & Jones,D.T. (2004) Predicting metal binding sites in low resolution structural models. *J. Mol. Biol.* **342**, 307-320.
4. Jones,D.T. & McGuffin,L.J. (2003) Assembling novel protein folds from super-secondary structural fragments. *Proteins: Structure, Function and Genetics* **53** (S6), 480-485.
5. Siew,N., Elofsson,A., Rychlewski,L. & Fischer,D. (2000) MaxSub: an automated measure for the assessment of protein structure prediction quality. Bioinformatics. **16**, 776-85.

**Keasar** - 283 models for 58 3D targets

## MESHI – a new object oriented package for molecular simulations

N. Kalisman, A. Levi and C. Keasar
*Department of Computer Science, Ben-Gurion Universiry, Israel*
keasar@cs.bgu.ac.il

MESHI is a novel software package that handles many aspects of molecular simulations. The motivation behind MESHI is twofold (1) to shorten the delay between the emergence of a novel idea (say, while one is doing the dishes) and the testing of its programmed manifestation. (2) to lower the "activation barrier" of the code, i.e. reduce the time it takes a new developer to start writing new modules. In order to achieve these goals, MESHI adheres to a strict Object Oriented Design (OOD) and emphases clear code, even at the expense of some computational efficiency. In CASP6 we tried to demonstrate that while still in a stage of development, MESHI has already crossed the critical point where useful molecular modeling is possible.

In practice, strict OOD implies that every aspect of molecular modeling is represented by a class. Thus, MESHI is equipped with classes for molecular elements (e.g. atoms and residues), geometric properties (e.g. distances and angles), energy-terms (e.g. hydrogen-bonds), optimization-algorithms (e.g. steepest-descent and LBFGS) and quite a few auxiliary classes (e.g. PDB formatted line). These classes serve as handy building blocks to MESHI applications like BEAUTIFY (the program we used for CASP6).

MESHI is written solely in Java, which is not the obvious language of choice for a computationally intensive program. Its interpreted nature is inherently slower than native binary code. Our experience is that java code is about two times slower than equivalent C/Fortran code. We believe however, that the most precious resource is the developer's time, as Moore's law does not apply to it. The strict object oriented nature of Java forces a highly modular program structure and helps in optimizing human effort. Further, Java's garbage-collection utility seems to remove a large family of bugs from our way. In practice, the performance loss of Java is much lower than twofold. By profiling one can easily identify the (typically few) bottlenecks where the program spends most of its time. These parts of the program may be written with emphasis on performance and/or compiled to a binary module.

Due of its low "activation barrier", MESHI is handy as an educational tool. In the last three years, students at the bioinformatics track of BGU did interesting and substantial projects within MESHI. The projects were defined in terms of interfaces and the students could focus in their specific tasks without diving into the code too deeply.

MESHI is free for academic use, and is available at:
http://www.cs.bgu.ac.il/~keasar/meshi

**KIAS** - 675 models for 64 3D / 64 DP / 64 RR

## Prediction of residue-residue contacts using correlated mutation and hydrophobic packing score

Mee Kyung Song, Keehyoung Joo and Jooyoung Lee*
*School of Computational Sciences, Korea Institute for Advanced Study*
*207-43 Cheongryangri-dong, Dongdaemun-gu, Seoul 130-722, Korea*
jlee@kias.re.kr

See methods section

## Tertiary structure prediction for comparative modeling, fold recognition and new fold targets in CASP6

Keehyoung Joo[1], Jejoong Yoo[1], Kyoungrim Lee[1], Hyung-Rae Kim[1], Seung-Yeon Kim[1], Mee Kyung Song[1], Ju-Beom Song[2], Sang Bub Lee[1,3], Sung Jong Lee[4], Jooyoung Lee[1*]
[1]*School of Computational Sciences, Korea Institute for Advanced Study*
[2]*Department of Chemistry, Kyungpook University, Korea;*
[3]*Department of Physics, Kyungpook University, Korea*
[4]*Department of Physics, Suwon University, Korea*
jlee@kias.re.kr

See methods section

**Luo** - 268 models for 54 3D targets

## Consistent scoring with AMBER/PB energy function

M.J. Hsieh and R. Luo
*Department of Molecular Biology and Biochemistry*
*University of California, Irvine, CA 92697*
rluo@uci.edu

See methods section
**MacCallum** - 128 models for 64 3D / 64 RR targets
**SBC** - 90 models for 64 3D targets
**DRIP-PRED** (serv) - 64 models for 64 DR targets
**GPCPRED** (serv) - 63 models for 63 RR targets

## Striped sheets, contact maps, disorder and model selection

R.M. MacCallum, B. Wallner and A. Elofsson
*Stockholm Bioinformatics Center, Stockholm University, Sweden.*
maccallr@sbc.su.se

In this poster, we will provide more in-depth data and figures for the methods already described in the following abstracts:

1. MacCallum & GPCPRED
Contact map prediction using PSI-BLAST profiles, self-organising maps and genetic programming.
2. MacCallum
Meta-server model selection using contact map-based scoring.
3. DRIP-PRED
Order/Disorder prediction using PSI-BLAST profiles and self-organising maps.

We also hope to show some preliminary analysis of our submitted predictions if target structures become available in time.

**MUMSSP** - 9 models for 2 3D targets

## How do the web facilities help predictors from head to toe of homology modeling?

M.R. Saberi, A. Razzazan, H. Ramezani and A. Baratian
*Medicinal Chemistry Division, School of Pharmacy, Mashhad University of Medical Sciences, Mashhad, Po. Box: 91775-1365, Iran*
saberimr@mums.ac.ir

See methods section
**Panther** - 55 models for 28 3D targets
**Panther2** (serv) - 48 models for 47 3D targets

## Prosite patterns for alignment validation and structural clusters as templates

Hao Wang, Robert W. Harrison
*Department of Computer Science, Georgia State University*

One recurring critical problem revealed in CASP has been the ability to model insertions and deletions in protein structure. Related to this is the inability of potential based modeling approaches to correct for minor sequence alignment errors. Threeapproaches were tested to see if they had potential to help overcome these issues. The first approach was to extend the molecular

mechanics potential by including a mean-force potential. The potential was chosen by defining a set of most common nodal or "eigenstructures" together with terms to represent the range of variation in the structure. These nodal structures effectively span the space of allowed and observed peptide conformations. The problem of modeling an insertion or deletion then becomes the problem of identifying the correct nodal structure. The nodal structures were chosen via K-nearest neighbors clustering to provide a uniform covering of the space of structures. The second approach was to add a switching hydrogen bond potential to help stabilize the backbone structure. This potential was implemented with a Morse function. Finally, sequence alignments were checked against a selected set of prosites patterns in order to validate the alignment under the assumption that similar structures would have a similar distribution of patterns. We also would expect that gain, loss, or shift in position of a pattern was indicative of a missalignment.

## **PROFESY** - 70 models for 14 3D targets

### Protein structure prediction method based on fragment assembly and conformational space annealing

Julian Lee[1], Seung-Yeon Kim[2] and Jooyoung Lee[2*]
[1]- *Dept. of Bioinformatics and Life Science, Soongsil University*,
[2] - *School of Computational Sciences, Korea Institute for Advanced Study*
jlee@kias.re.kr

See methods section

## **Rokko** - 228 models for 64 3D targets

### *De novo* structure prediction by SimFold: benchmark test and comparison with Rosetta

Y. Fujitsuka[1*], G. Chikenji[2*], S.J. Park[3], W. Jin[2] N. Koga[1],
T. Furuta[2], and S. Takada[1,2]
[1] – *Grad School, Sci & Tech Kobe Univ*, [2] – *Faculty of Sci, Kobe Univ*, [3]- *Interdisciplinary Grad School of Sci & Eng, Tokyo Inst Tech*
stakada@kobe-u.ac.jp

We have developed a method for *de novo* protein structure prediction and compared its performance with Rosseta[1]. In our approach, first, we prepare the fragment candidates of every 10 residues for each position of target proteins. Then, we perform fragment assembly simulation with reversible replacement[2] using our in-house developed energy function, SimFold[3,4]. We carried out a small scale benchmark test on a set of proteins selected in Baker's paper[5]. For comparison, we also use Rosetta *ab initio* software with default parameters on the same set. Relative performance depends on proteins; some are predicted better by Rosetta, others by our approach, and the rest predictions are equivalent. Overall comparison indicated that the method developed here performs slightly better, on average, than Rosetta. We (Team ROKKO) also applied our method to all the CASP6 targets that possibly have "new folds". In particular, we succeeded in predicting the correct fold of T0198 with ~8Å RMSD accuracy.

Our strategy consists of following four elements: 1) generation of fragment candidates, 2) designing energy functions, 3) conformational sampling by the fragment assembly (FA) and 4) selecting models.

1) Generation of fragment candidates: We used different methods in two different purposes. a) For the benchmark test, fragment candidate are prepared by Rosetta Fragment Selection software to make comparison as fair as possible. b) For the CASP6 query, methods are described in the Method abstract of team ROKKO.

2) SimFold, the energy function[3,4]: The protein is represented by a coarse-grained model, in which side chain atoms are replaced by a center of interactions. The interaction potential which we call SimFold contains van der Waals interaction, secondary structure propensity, the hydrogen bond interaction, the hydrophobic interaction and the pair-wise interaction. Some more details are found in ROKKO's method abstract.

3) Fragment assembly: For conformational sampling, we use a variant of fragment assembly (FA) method called "reversible FA method" which we have recently developed (an earlier version in ref.3). Our FA is different from what has been developed by Baker's group[1]. The most important difference between conventional FA and ours is that the conventional FA protocol does not fulfill the detailed balance condition, but our algorithm does. This enables us implementing powerful generalized ensemble methods such as replica-exchange and multi-canonical ensemble methods. The latter was indeed used in the CASP6. For benchmark comparison, simple simulated annealing is used.

4) Model selection: We carried out cluster analysis for the structures generated by FA simulations. For the benchmark test proteins, the 5 models are chosen from the 5 largest clusters automatically. For the CASP6 targets, if whole-length structures are not well clustered, the substructures are attempted to be clustered. Then, the representatives of larger clusters are chosen as models based on human inspection.


1. Simons,K.T., Kooperberg,C., Huang,E. & Baker,D. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* **268**, 209-225
2. Chikenji,G., Fujitsuka,Y., and Takada,S. (2003) A reversible fragment assembly method for *de novo* protein structure prediction. *J.Chem.Phys.* **119**, 6895-6903
3. Takada,S. (2001) Protein folding simulation with solvent-induced force field: Folding pathway ensemble of three-helix-bundle. *Proteins* **42**, 85-98.
4. Fujitsuka,Y., Takada,S., Luthey-Schulten,Z.A., and Wolynes,P.G. (2004) Optimizing physical energy functions for protein folding, Proteins **54**, 88-103.
5. Simons,K.T., Strauss,C., & Baker,D. (2001) Prospects for *ab initio* protein structural genomics, Proc. Natl. Acad. Sci. USA. **306** 1191-1199.


## **rost_PROFcon** (serv) - 64 models for 64 RR targets

### PROFcon - a new neural network-based contact predictor


M. Punta[1,2] and B. Rost[1,2,3]

*[1] -CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY 10032, USA, [2] - Columbia University Center for Computational Biology and Bioinformatics (C2B2), New York, NY 10032, USA,, [3] - NorthEast Structural Genomics Consortium (NESG), Department of Biochemistry and Molecular Biophysics, Columbia University, New York, USA*

punta@cubic.bioc.columbia.edu


See methods section

## **SAM-T04-hand** - 375 models for 64 3D / 56 RR targets

### Human interaction with undertaker for the structure prediction of targets T0212 and T0198


Martina Koeva and Kevin Karplus

*University of California, Santa Cruz*

martina@soe.ucsc.edu


We present two examples – T0212 and T0198 – of human intervention in the protein structure prediction process through our interaction with the "undertaker" program. Preliminary analysis of the results for these two targets allows us to gain some understanding of the abilities and limitations of the program, as well as to assess the human-added value to the quality of the predictions.

Target T0212 consisted of approximately 126 residues and was annotated as protein SOR45 from S.oneidensis. We used a fully automated method, which involved the use of SAM-T04, SAM-T2K and undertaker[1], to generate an initial 3D model for this target. Our initial alignment results did not suggest any obvious templates (for comparative modeling) or folds (for fold recognition). Based on the structural neighbors of our initial models and some of the sequence alignments, we decided to pursue a jelly-roll like topology. Our secondary structure predictions suggested that if we modeled T0212 as a jellyroll, our models were going to have either an extra strand, or a missing strand. We used undertaker to pursue both possibilities. The comparisons of our results with the correct structure (PDB: 1tza) indicate that our top submitted model, which represented the equivalence class of the "jelly-roll with a missing strand" models scored the best from all of our submitted models with a GDT score of 30.645%.

Target T0198, which corresponded to protein 1170B from Thermotoga maritima had a sequence of length 235 amino acids. The initial sequence alignments and secondary structure predictions suggested a helical up-and-down bundle fold, which our initial 3D model generated by undertaker did not reflect. We decided to pursue two different possible folds: an alpha-helical sandwich, based on some of the structural neighbors of T0198, and a helical bundle. We did not manage to use undertaker to successfully bundle the predicted helices. We could not find a bundling pattern that allowed us to make undertaker pack tightly the helices against each other, while exhibiting the appropriate exposure/burial patterns. Undertaker seemed to favor mostly alpha-

helical sandwich models. The comparison between our submitted models and the correct structure (PDB: 1sum) has shown poor results and not much improvement over the initial automatically generated model. Our best model, which was not submitted, showed a GDT score of 21% and 12.54 Ang. RMSD.


1. Karplus,K., Karchin,R., Draper,J., Casper,J. Mandel-Gutfreund,Y., Diekhans,M., and Hughey,R. (2003) Combining local-structure, fold-recognition, and new-fold methods for protein structure prediction. *Proteins* **53** S6, 491-496.


## Residue-residue contact prediction using mutual information and neural networks


George Shackelford, Kevin Karplus
*University of California, Santa Cruz*
karplus@soe.ucsc.edu


We present a neural network predictor of residue-residue contacts that uses statistical analysis of mutual information and local property values as inputs. The results improves on earlier efforts[1,2.]

Two problems with earlier efforts in using mutual information result from small sample size and biased sampling due to over-representation of sub-family sequences in the alignment. We show ways to deal with both these problems by two statistical methods for correction of small samples and an aggressive thinning of the sequences.

We use SAM-T04 to get the alignments[3]. For each pair we randomize the contingency table while holding fixed the marginal sums, and build a histogram of the mutual information for each randomization. We use this histogram to adjust for small sample sizes in two ways. The first corrects for mutual information based on chance by subtracting the mean of the histogram from the raw mutual information to give a corrected mutual information. The second takes the histogram and fits a gamma distribution on it. We use that distribution to calculate an e-value. Both of these values show a significant improvement over raw mutual information.

We compensate for the bias of over-represented sequences by thinning the sequences to a series of subsets with increasing dissimilarity between the sequences. We find that thinning in general improved results and thinning to 35% sequence similarity between all sequences provides the best results in balancing between the bias and sample size.

Finally we are able to improve on these predictions by using these as part of the inputs to a neural network. The network consists of 280 inputs consisting of sequence length, separation, corrected mutual information and e-values for different thinnings, distributions of both residues including neighboring residues, and burial and secondary structure predictions. The network's single output is the probability value that there is a contact between the respective residues. The results of preliminary testing suggest a significant improvement over previous predictors.

The predictions were available as constraints for the "undertaker" program here at UC, Santa Cruz. There are no current results to show whether or not those constraints improved the tertiary structure predictions.


1. Gobel,U., Sander,C., Schneider,R., Valencia,A. (1994) Correleated mutations and residue contacts in proteins. *Proteins* **18**, 309-317.
2. Fariselli,P., Olmea,O., Valencia,A., Casadio,R. (2001) Prediction of contact maps with neural networks and correlated mutations. *Protein Engineering* **14** (11), 835-843.
3. Karplus,K., Karchin,R., Draper,J., Casper,J. Mandel-Gutfreund,Y., Diekhans,M., and Hughey,R. (2003) Combining local-structure, fold-recognition, and new-fold methods for protein structure prediction. *Proteins* **53** (S6), 491-496.


# **SBC** - 90 models for 64 3D targets

## A study of different profile-profile alignment methods


Tomas Ohlson, Björn Wallner and Arne Elofsson
*Stockholm Bioinformatics Center*
tomasoh@sbc.su.se


It has been demonstrated that methods using multiple sequences, i.e. evolutionary information, are superior to methods that only use single sequences[1] and more recently that methods that use evolutionary information for both the query and target sequences are even more efficient[2] when it comes to detecting homologous proteins. One method to include this information is by the use of profile-profile alignments, where a profile from the query protein is compared with the profiles from the target proteins.

Profile-profile alignments can be implemented in several fundamentally different ways. The similarity between two positions can be calculated using a dot-product, a probabilistic model or an information theoretical measure. In addition, information about the background frequency of amino acids can be used.

In this study[3] we present a large scale comparison of different profile-profile alignment methods. We show that the profile-profile methods perform at least 30% better than standard sequence-profile methods both in their ability to recognize superfamily related proteins and in the quality of the obtained alignments. The main reason behind the improvement is most likely that the profile-profile scoring methods are better at distinguishing evolutionary related positions from non-related positions.


1. Lindahl,E., Elofsson,A. (2000). Identification of related proteins on family, superfamily and fold level. *J. Mol. Biol.* **295**, *613-625.*
2. Wallner,B., Fang,H., Ohlson,T., Frey-Skött,J., Elofsson,A. (2004). Using evolutionary information for the query and target improves fold recognition. *Proteins* **54**, *342-350.*
3. Ohlson,T., Wallner,B., Elofsson,A. (2004). Profile-profile methods provide improved fold-recognition: A study of different profile-profile alignment methods. *Proteins* **57,** 188-197**.**


## Benchmark of different homology modeling packages


B. Wallner and A. Elofsson
*Stockholm Bioinformatics Center, Stockholm University*
bjorn@sbc.su.se


In this study, we have used alignments between protein domains belonging to the same SCOP family, with sequence identity ranging from 30%-100%, as an input to six different homology modelling programs, Modeller[1], SegMod/ENCAD[2], SWISS-MODEL[3], 3D-JIGSAW[4], Builder[5,6] and *nest*[7] within the JACKAL modeling package[8]. As a further reference SCWRL3[9] and (also within the Jackal modeling package) were used to build side-chain from simple backbone models. The overall quality and stereochemistry of the resulting models were analyzed.

In general there is not a huge difference between the different methods. But looking at the details there are differences. The differences are most pronounced for the side-chain prediction, were there is clearly room for improvement. For backbone dihedrals all methods perform equal except SegMod/ENCAD which has 10 percentage points lower fraction of phi/psi dihedrals correct. However these models have good stereochemistry which indicates that it is difficult to get both the correct stereochemistry and correct backbone dihedrals.

It has been shown in many studies and also at CASP that a model very seldom is more close to the native structure than the template it was build on. This is also true for most cases in this benchmark. However some methods like *nest* very rarely makes the model worse, resulting in higher fraction of models that get better compared to models that get worse (5% vs 2.5%).

Overall SegMod/ENCAD, Modeller and nest produce a higher number of acceptable models compared to the other methods.


1. Sali,A. and Blundell,T.L. (1993). Comparative modelling by statisfaction of spatial restraints. *J. Mol. Biol.* **234**(3), 779-815.
2. Levitt,M. (1992). Accurate modeling of protein conformation by automatic segment matching. *J. Mol. Biol.* **226**(2), 507-533.
3. Schwede,T., Kopp,J., Guex,N. and MC Peitsch (2004). SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res.* **31**(7), 3381-3385.
4. Bates,P.A., Kelley,L.A., MacCallum,R.M. and Sternberg,M.J. (2001). Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. *Proteins* **Suppl 5**, 39-46.
5. Koehl,P. and Delarue,M. (1995). A self consistent mean field approach to simultaneous gap closure and side-chain positioning in homology modelling. (1995). *Nat. Struct. Biol.*, **2**(2), 163-170.
6. Koehl,P. and Delarue,M. (1994). Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J. Mol. Biol.* **239**(2):249-275.
7. Petrey,D., Xiang,Z., Tang,C.L., Xie,L., Gimpelev,M., Mitros,T., Soto,C.S., Goldsmith-Fischman,S., Kernytsky,A., Schlessinger,A., Koh,I.Y., Alexov,E. and Honig,B. (2003). Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology

modeling. *Proteins* **53 Suppl 6**:430-435.
8. Xiang,J.Z. (2003) Jackal: A protein structure modeling package. http://trantor.bioc.columbia.edu/programs/jackal/index.html.
9. Canutescu,A.A., Shelenkov,A.A. and Dunbrack,R.L Jr. (2003). A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci.* **12**(9), 2001-2014.

## **Softberry** - 122 models for 63 3D / 59 DR targets

### SoftPM: Softberry tools for protein structure modeling

V. Solovyev [1,2], D. Affonnikov[2], A. Bachinsky[2], I. Titov[2], V. Ivanisenko[2] and Y. Vorobjev[2]
[1]- *Department of Computer Science, Royal Holloway, University of London, Egham, Surrey TW20 0EX,UK;*
[2]-*Softberry Inc., 116 Radio Circle, Suite 400; Mount Kisco, NY 10549, USA*
victor@cs.rhul.ac.uk

See methods section

## **Tramontano** – organizer, no predictions

### PMDB: a new freely accessible database of protein structure models

P. D'Onorio De Meo, D. Cozzetto, V. Zafiropoulos, C. Valeriani, T. Castrignan, and A. Tramontano

The Protein Models DataBase (http://sandokan.caspur.it/PMDB/) collects three-dimensional protein models obtained by any structure prediction method and labelled with a reliability value. The system allows users both to contribute new models and to search for existing ones. The database currently stores all models submitted to the last edition of the CASP experiment.

## **VENCLOVAS** - 25 models for 25 3D targets

### PSI-BLAST-ISS: an intermediate sequence search tool for estimation of position-specific alignment reliability

M. Margelevičius and Č. Venclovas
*Institute of Biotechnology, Graičiūno 8, Vilnius, Lithuania*
venclovas@ibt.lt

The Intermediate Sequence Search (PSI-BLAST-ISS) tool is designed to assess the region-specific alignment reliability between two protein sequences (target and template). The main idea of the algorithm is to initiate additional PSI-BLAST[1] searches against the non-redundant sequence database for a set of sequences that are related both to the target and to the template[2]. The position-specific reliability of the alignment between the target and the template is then assessed by merging alignment data obtained from intermediate sequence searches and analyzing alignment convergence.

Algorithm

The whole ISS procedure may be described as the following steps: (1) identification of multiple sequences related to both target and template sequences, (2) creation of a representative set from these sequences by filtering out close homologs, (3) generation of multiple sequence alignments for all sequences from this representative set by searching sequence database, containing both target and template sequences, (4) retention of all instances of significant matches between the target and the template from multiple alignments obtained in step 3, (5) merging all of significant target-template alignments by taking one of the sequences (either the target or the template) as a frame of reference. Optionally, the procedure can include creation of the consensus template sequence derived from the final merged target-template alignment. Using this option, the position specific reliability for multiple target-template alignments can be contrasted simultaneously.

Implementation

PSI-BLAST-ISS is a collection of fairly independent modules linked together using Perl. As an input, PSI-BLAST-ISS takes the target sequence, which is searched against the non-redundant sequence database to collect intermediate sequences. The set of intermediate sequences is currently filtered by CD-HIT[3], the sequence clusterization program. Each of the intermediate sequences is used to generate sequence profiles in a form of PSI-BLAST checkpoint file by running a user-defined number of PSI-BLAST iterations. The resulting

checkpoint files are then used to restart PSI-BLAST searches in a sequence database that has to include sequences of both proteins of interest (target and template). In a common situation, when the template represents a structural template intended for use in comparative modeling, such database may be derived by simply appending the target sequence to the PDB sequence database that already contains the template sequence. After the processing and merging obtained target-template alignment variants the final result is a multiple sequence alignment, where the reference sequence (say the target) is aligned with multiple instances of the second sequence (template) according to different alignment variants.


1. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
2. Venclovas, Č. (2001). Comparative modeling of CASP4 target proteins: combining results of sequence search with three-dimensional structure assessment. *Proteins* ***Suppl 5***, 47-54.
3. Li,W., Jaroszewski,L., Godzik,A. (2001) Clustering of highly homologous sequences to reduce the size of large protein database. *Bioinformatics* **17**, 282-283.


## **Wymore** - 32 models for 19 3D targets

### Comparative modeling using alternative alignments and statistical potentials


Adam Marko, Stuart Pomerantz, Troy Wymore
*Biomedical Initiative Group, Pittsburgh Supercomputing Center, Pittsburgh, PA*
wymore@psc.edu


For even moderately difficult comparative modeling projects, there is often variable regions for which the alignment between target and template is highly arbitrary and hence structures generated through such an alignment can have significant errors. In an effort to overcome these errors, we have developed a protein structure prediction pipeline that is currently applicable for these comparative modeling targets. This pipeline consisted of 1) generating hundreds of alternative alignments between target and template 2) using these alignments to generate structures 3) scoring these structures with a statistical potential and 4) visually examining lowest energy structures in an effort to pick the one closest to native. Programs were written in Perl to enable the flow between modeling programs. Our goal for this part of our modeling strategy was to demonstrate improvement in our comparative models over those constructed from a T-coffee[1] alignment.

Template structures were identified by performing a BLAST[2] search through the non-redundant database, building profiles from related sequences through the MEME[3] program and using those profiles to search through the PDB using the MAST[4] program. We constructed 100-500 alternative alignments between template and target using the program probA[5]. This program uses a probabilistic backtracking procedure that generates ensembles of suboptimal alignments with correct statistical weights. This ensemble of alignments was used to build structures using MODELLER version 6.2[6]. The structures were then ranked using ProsaII[8]. For some targets, we attempted to distinguish between favorable "ProsaII" models with an all-atom molecular mechanical potential coupled to a Generalized Born implicit solvent model. This presentation will describe the 1) the ability of the ProsaII program to identify structures closest to native from an ensemble and 2) the improvements in alignment quality and native contacts generated through the use of this pipeline versus constructing a model from a T-coffee multiple sequence alignment.


1. Notredame, C., Higgins, D., Heringa, J. (2000) T-Coffee: A novel method for multiple sequence alignments. *J. Mol. Biol.*, **302**, 205-217.
2. Altschul,S.F., Gish, W., Miller, W., Meyers, E. W., Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410.
3. Bailey, T. L., Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. 2nd Int. Conf. Intelligent Sys. Mol. Biol.* AAAI Press, 28-34.
4. Bailey, T. L., Gribskov, M. (1998) Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* **14**, 48-54.
5. Kelley, L. A., MacCallum, R. M., Sternberg, M. J. E. (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.* **299**, 499-520.
6. Muckstein, U., Hofacker, I. L., Stadler, P. F. (2002) Stochastic pairwise alignments. *Bioinformatics*, **18**, S153-S160.
7. Sali, A., Blundell, T. L. (1993) Comparative Protein Modeling by Satisfaction of Spatial Restraints. *J. Mol. Biol.*, **234**,779-815.
8. Sippl, M. J. (1993) Recognition of Errors in Three-Dimensional Structures of Proteins. *PROTEINS: Struct. Func. Gen.* **17**,355-362.

## Loop modeling using the Multi-scale Modeling Tools for Structural Biology (MMTSB) toolset

Troy Wymore and Adam Marko
*Biomedical Initiative Group, Pittsburgh Supercomputing Center, Pittsburgh, PA*
wymore@psc.edu

Our group has developed a comparative modeling pipeline that attempts to generate a model that would correspond to the structural alignment as much as possible. Yet even if the optimal target-template alignment is generated and identified, there are often insertions or highly variable regions that will contain significant errors. In these instances, one must resort to other methods such as physics-based simulations to obtain a structure closer to the native. This presentation will describe loop refinement efforts using the MMTSB[1] Toolset during the prediction season and long molecular dynamics simulations performed afterward.

For six highly variable regions ranging in size from 5-16 residues, we performed lattice-based replica exchange simulations using MONSSTER[2] through the MMTSB toolset for enhanced sampling of conformational space. Restraints were placed on the rest of the structure. The lowest temperature replicas from the final rounds of simulation (typically the last 100-1000 structures) were rebuilt to complete all-atom models. These structures were minimized with the all-atom force field in CHARMM with a distance dependant dielectric function. Energies for these structures were then evaluated more accurately with the same force field but coupled with a Generalized Born implicit solvent model. The loop structures were clustered according to distance RMSD. And finally a model was chosen with the lowest energy in the cluster with the lowest average energy.

1. Feig,M., Karanicolas,J., Brooks III,C.L.B. (2004) MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology. *J. Mol. Graph. Model.* **22**, 377-395.
2. Skolnick,J., Kolinski,A., Ortiz,A.R. (1997) MONSSTER: a method for folding globular proteins with a small number of distance restraints. *J. Mol. Biol.* **265**, 217-241.

**YASARA** - 28 models for 9 3D targets

## The last mile of the protein folding problem – a pilgrim's staff and skid-proof boots

E. Krieger, S.B. Nabuurs, C.A.E.M. Spronk and G. Vriend
*CMBI, Center for Molecular and Biomolecular Informatics, Radboud University Nijmegen, the Netherlands*
Elmar.Krieger@cmbi.ru.nl, www.YASARA.org

Today's energy functions are not able yet to distinguish reliably between correct and almost correct protein models. Improving these near-native models is currently a major bottle-neck in homology modeling or experimental structure determination at low resolution. Increasingly accurate energy functions are required to walk along the '***last mile of the protein folding problem***', for example during a molecular dynamics simulation.

Here we provide a pilgrim's staff: self-parameterizing force fields[1], that were obtained from the AMBER force field[2] by simulating complete protein crystals and iteratively adjusting the force field parameters to minimize the damage done to the known structures[3]. The resulting YAMBER and YASARA force fields are then used to run accurate simulations of homology models in aqueous solution.

Additional skid-proof boots are needed to avoid a common pitfall: even with an ideal force field, homology models cannot be expected to always approach the native conformation directly. That's why we run 100 simulations in parallel[4] and then use a sophisticated scoring function based on WHAT IF checks[5] and YASARA energies to pick out the pearls.

Even models very close to the native structure can be improved: The closest template for Target 231 was an NMR structure with 80% sequence identity and 0.96 Å Cα RMSD (excl. one long flexible surface loop). During the refinement, this RMSD could be reduced to 0.79 Å (Model 1).

More information is available at www.yasara.org and www.cmbi.ru.nl/whatif

1. Krieger,E., Koraimann,G. & Vriend,G. (2002) Increasing the precision of comparative models with YASARA NOVA - a self-parameterizing force field. *Proteins* **47**, 393-402.
2. Wang,J., Cieplak,P. & Kollman,P.A. (2000) How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J.Comp.Chem.* **21**, 1049-1074.
3. Krieger,E., Darden,T., Nabuurs,S.B., Finkelstein,A. & Vriend,G. (2004) Making optimal use of empirical energy functions: force field parametrization in crystal space. *Proteins* **in press**.
4. Krieger,E. & Vriend,G. (2002) Models@Home: distributed computing in bioinformatics using a screensaver based approach. *Bioinformatics* **18**, 315-318.
5. Hooft,R.W.W., Vriend,G., Sander,C. & Abola,E.E. (1996) Errors in protein structures. *Nature* **381**, 272-272.

## **Accelrys** - 27 models for 16 3D / 1 FN targets

### Modeling, simulations and high-throughput functional annotation using Discovery Studio Modeling and GeneAtlas

D. Haley-Vicente
*Accelrys Inc., 9685 Scranton Rd., San Diego, CA 92121*
dhv@accelrys.com

A plethora of methodologies have been utilized for CASP6 homology model predictions. We determined the protein models based on a combination high-throughput bioinformatics, modeling and simulations algorithms in Discovery Studio® (DS) Modeling (Accelrys, Inc)[1]. As part of DS Modeling, an automated, high-throughput functional annotation pipeline program called DS GeneAtlas[2] was used to predict the majority of templates and provide initial alignments and models for each target. The DS GeneAtlas pipeline incorporates sequence similarity detection (e.g. PSI-BLAST), domain analysis (e.g. PFAM), homology modeling (e.g. MODELER), model evaluation (e.g. Profiles-3D), fold recognition (e.g. SeqFold), and 3D active site annotation (e.g. CSC[3] 3D-motif searching) methods.

Both DS Modeling and DS GeneAtlas will be demonstrated at the CASP6 conference in Gaeta, Italy (December 2004). The demonstration will show advanced *in silico* high throughput bioinformatics, functional annotation, protein homolog modeling and 3D annotation techniques to study genomes and proteomes. The software demonstration includes using our Discovery Studio software to analyze the West Nile Virus (WNV) genome[4]. The software demonstration will show that the DS GeneAtlas pipeline can be used to produce reliable structural and functional annotation of the WNV capsid, envelope, NS1, NS3, and NS5 proteins. Functional annotation for these proteins reveal information regarding their predicted transmembrane region, structure, function and binding site(s). The 3D homology model of the proteins can then be used as the biological target for lead finding experiments that include a combination of docking and *de novo* design.

1. Discovery Studio Modeling (http://www.accelrys.com/dstudio/ds_modeling/) Accelrys Inc.
2. Kitson, et al. (2002) Functional annotation of proteomic sequences based on consensus of sequence and structural analysis. *Briefings in Bioinformatics* **3**, 1-13.
3. Milik, et al. (2003) Common Structural Cliques: a tool for protein structure and function analysis. *Protein Engineering* **16,** 1-10.
4. Quinn, Fisher and Haley-Vicente (2004) From Gene to Function: *In Silico* Warfare on the West Nile Virus (http://www.accelrys.com/cases/west_nile_virus.pdf)

## **HHpred.2** (serv) - 310 models for 62 3D targets
## **HHpred.3** (serv) - 309 models for 62 3D targets

### HHpred web server for distant homology detection and structure prediction

J. S Söding, A. Biegert, A. Lupas
*Dept for Protein Evolution,*
*Max-Planck-Institute for Developmental Biology, Tübingen, Germany*
johannes.soeding@tuebingen.mpg.de

HHpred is a server for the detection of distant homologs that can also be used for structure prediction. The user can paste either a query sequence or a whole alignment. HHpred then performs a specified number of PSI-BLAST iterations (between 0 and 8) against the non-redundant database and predicts secondary structure with PSIPRED. An HMM is generated for the query alignment and the query HMM is compared with a user-selected database of HMMs. At the moment, HHpred allows searching Pfam, SMART, and SCOP. Inclusion of DALI and a daily updated version of the PDB protein data bank are planned. The user can use local or semi-global HMM-HMM alignment for the search and can choose whether to include secondary structure scoring. (If the 2D structure of the database sequences is not known, predicted secondary structure is used instead.)

HHpred returns a list of best matches together with the query-template alignments in an easily readable format. The alignments include the secondary structures and the consensus sequences of query and template, as well as a line showing the match quality of each pair of HMM columns. Furthermore, the user can chose to include up to ten representative homologs of the query and template in the alignments and he may color residues by biochemical similiarity. Hits are linked to Pfam, SCOP, SMART, and/or the PDB.

The user may view the query and template alignments from which the HMMs were calculated. He may edit the query alignment and resubmit the corrected alignment to HHsearch. This ensures full flexibility for interactive use. The user may also choose to generate a query-template alignment (with one or multiple templates) as input to homology modelling programs. At the moment, FASTA and PIR format is supported. Alternatively, an unrefined 3D structure model in pdb format containing only the $C_\alpha$ atoms can be generated by mapping the coordinates of the template to the query residues in accordance with the query-template alignment found.

HHpred is very fast: a search against the SCOP50 database (~6700 domains) with a 200-residue query takes about one minute, plus the time to build an alignment with PSI-BLAST. To make the server easy to use we have added a help facility that explains the input parameters as well as how to interpret the search output. The server distributes jobs to a small compute cluster which will be extended as required.

We plan to extend HHpred together into a flexible structure and function prediction pipeline for interactive use. We welcome suggestions for improvement or further development. You can access HHpred at http://protevo.eb.tuebingen.mpg.de/toolkit/index.php?view=hhpred.

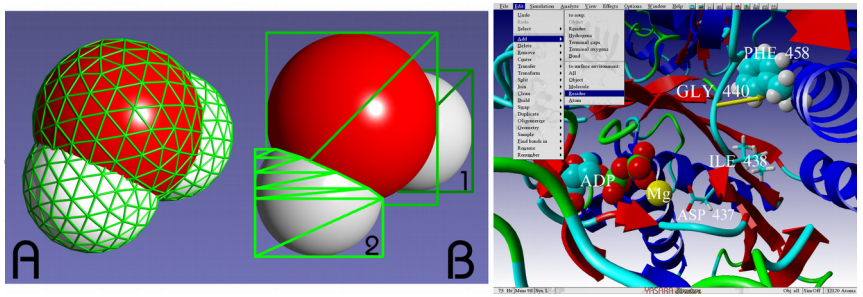
1. Söding,J. (2004) Protein homology detection by HMM-HMM comparison. Submitted to *Bioinformatics*.

**YASARA** - 28 models for 9 3D targets

## YASARA – Molecular graphics, -modeling and -simulation

Elmar Krieger
*CMBI, Center for Molecular and Biomolecular Informatics,*
*Radboud University Nijmegen, the Netherlands*
Elmar.Krieger@cmbi.ru.nl, www.YASARA.org

Because our brain prefers images over numbers, progress in the natural sciences is coupled with the ability to display and investigate molecules on a computer screen. Nowadays, computers are equipped with graphics processing units (GPUs), that heavily accelerate the display of three-dimensional models. Molecular visualization algorithms run into an unexpected problem, however: GPUs are highly optimized for drawing triangles, while atoms are typically shown as plain spheres. Programmers are thus forced to join ~320 or more triangles to display one single atom (Figure A). For large biomolecular systems with tens of thousands of atoms, this approach becomes prohibitively slow. Here I describe a novel way of drawing molecules, that requires a minimum number of two triangles per atom. These flat triangles have a precalculated image of a sphere attached which creates the illusion of depth. When compared with the classical approach, the novel method is up to 35 times faster, especially when visualizing large structures like the ribosome or virus capsides. An implementation of the algorithm is freely available as part of YASARA, a molecular graphics, modeling and simulation program for Linux and Windows, with support for structure analysis and prediction, interactive real-time simulations using classic and newly developed force fields, molecular animations, interactive tutorials, multimedia presentations, Python plugins and Yanaconda macros at www.yasara.org.



**I**



**J**



**K**



**L**



**M**



**N**



**O**

**X**



**Y**



**Z**