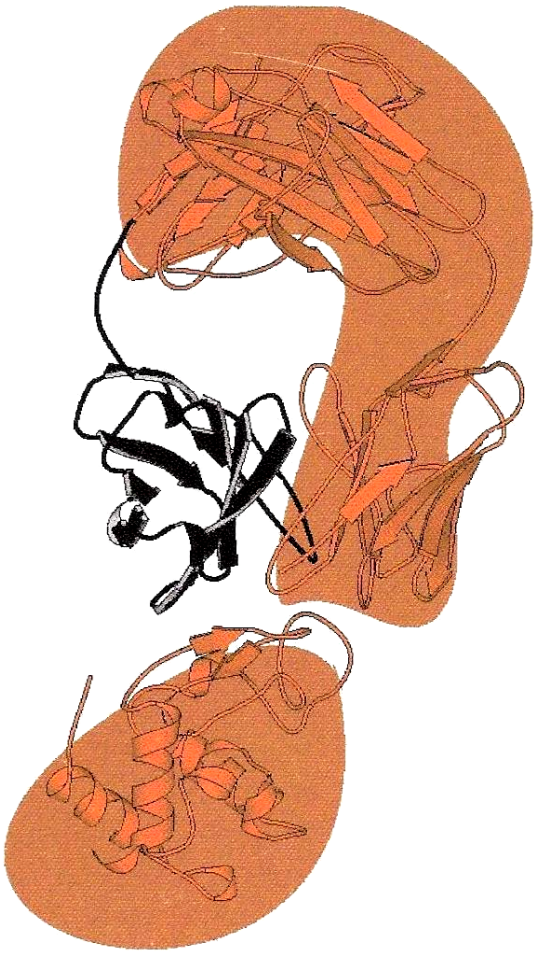
Critical Assessment of Techniques for Protein Structure Prediction

Eighth Meeting

C A S P 8



CAGLIARI, SARDINIA, ITALY
DECEMBER 3-7, 2008

ABSTRACTS

8th Community Wide Experiment on the
Critical Assessment of Techniques for Protein Structure Prediction

Cagliari, Sardinia, Italy
December 3-7, 2008

CASP8 Organizers:

John Moult	CASP President, CARB, University of Maryland, USA
Krzysztof Fidelis	University of California, Davis, USA
Andriy Kryshtafovych	University of California, Davis, USA
Burkhard Rost	Columbia University, New York, USA
Anna Tramontano	University of Rome, Italy

Funded by:
National Institutes of Health, NLM and NIGMS
BioSapiens Network of Excellence
EMBO

3D-JIGSAW_V3
3D-JIGSAW_AEP
Bates_BMM

Template and fragment mixing using a genetic algorithm

M.N. Offman, R.A.G. Chaleil, I. Moal and P.A. Bates
Cancer Research UK London Research Institute
paul.bates@cancer.org.uk

Our objective is to blend models created by several different means, in an attempt to combine the good quality regions from each into a final, more refined, model. We have developed a number of refinement operators (the move-set) to search restricted regions of conformational space. These operators are used in the context of a genetic algorithm (GA) that reshuffles and repacks structural components at both a finer local and a coarser global level1-2.

For CASP8 we entered two fully automated servers, 3D-JIGSAW_V31 and 3D-JIGSAW_AEP2 both of which employ this GA approach. Potential templates and fragments are first identified using the HHpred software package3. All templates are modelled to the target sequence using the sidechain replacement program SCWRL34. Insertions and deletions are modelled by our in-house loop modelling and closure method2, a complex procedure including a modified version of the cyclic coordinate descent algorithm5. The initial population of models is ranked with our course energy scoring function2 before being fed into several complete rounds of GA optimisation. These rounds of GA optimisation employ the move-set and model selection as previously described[1-2].

In the GA approach used, each round normally consists of a diversification (sampling) and an intensification (ranking) step. 3D-JIGSAW_AEP differs from 3D-JIGSAW_V3 in that the former uses Alternating Evolutionary Pressure (AEP), a new method to increase and improve diversification. In general, GAs and other similar conformational search algorithms can suffer from the problem that they tend to stay within local minima instead of exploring further afield.

Therefore, in our AEP approach, a number of consecutive diversification steps are allowed between each intensification step. In these non-scored rounds, the population grows linearly and the structures in the ensemble are allowed to sample energetically unfavourable intermediate states. Although energy evaluation is not applied, to ensure reasonable sampling, basic protein health checks associated with the operators are still in place.

In our manual submission protocol, all server models were downloaded from the prediction center webpage and used as the input population to 3D-JIGSAW_V3 in the upload mode. After recombination the highest scoring models, always at least slightly different from all initial input models, were submitted.

Our own preliminary analysis suggests that between our fully automatic servers, 3D-JIGSAW_V3 performs better for the easier targets and 3D-JIGSAW_AEP for the more difficult ones. This is perhaps not too surprising since 3D-JIGSAW_AEP is able to search more conformational space. However, It is encouraging that our energy function can score some of the models from the AEP approach well. On the negative side, the AEP method still has a tendency to move away from the target minima into deep, false positive, energy basins. Baring this in mind, we plan to merge both servers and apply the techniques depending on the targets' modelling difficulty levels. Furthermore, we are currently trying to reduce entry into false positive energy basins by adding protein family specific knowledge thereby locating more optimal regions for performing protein model crossover and mutation events.

1. Offman M.N., Fitzjohn P.W. & Bates P.A. (2006) Developing a move-set for protein model refinement. Bioinformatics. 22, 1838-1845.
2. Offman M.N., Tournier, A.L. & Bates, P.A. (2008) Alternating evolutionary pressure in a genetic algorithm facilitates protein model selection. BMC Struct. Biol. 8:34.
3. Soding, J., Biegert, A. & Lupas A.N. (2005) The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res.,33, W244-W248.
4. 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.
5. Canutescu, A.A. & Dunbrack R.L.Jr. (2003) Cyclic coordinate descent: a robotics algorithm for protein loop closure. Protein Sci.,12, 963-972

3DShot1
3DShotMQ
3DShot2

Novel, Meta-Approach based techniques for protein structure prediction

T. Seth and D. Fischer
tseth@cse.buffalo.edu, df33@cse.buffalo.edu

In the eighth Critical Assessment of techniques for protein structure prediction (CASP 8) experiment, we participated as 3DShot1, 3DShotMQ in the "Human-Server" category and, 3DShot2 in the "Server-Only" category. All the three of our methods generate automated predictions using three different meta-selection and assembly techniques.

3DShot1:
3DShot1 incorporates the following steps:
1) Input Selection: 3DShot1 evaluates all the "TS1" models submitted by the CASP7 server groups and extracts a subset 8-15 models for generating a hybrid model. The subset of models is determined based on a clustering scheme that takes into consideration different factors like the model quality, alignment length etc.
2) Input Refinement: Each of the selected input models is refined (Beautify) in order to remove collisions. Models from highly diverse clusters are further assessed for quality using MQAP.
3) Assembly: For each residue position, a set of its spatially closest residues is determined. These sets are then scored based on various sequence and structure-based properties and the best scoring sets are assembled together in a controlled environment to generate a hybrid prediction.
4) Assembly Selection and Refinement: Each of the final assembled models is ranked based on a composite score derived as a function of the structural similarity of the assembled model with the other models and the 3D-1D scores. The highest-ranking assembled model is refined and is reported as the final result.

3DShotMQ:
3DShotMQ incorporates the same initial steps as 3DShot1 but generates hybrid models using a different set of models and a different version of the assembly algorithm.
1) Input Selection and Refinement: Fifteen input models are obtained using a new local meta-mqap method and are refined using Modeller.
2) Assembly: The assembly is carried out using a slightly different version of the assembly algorithm used for 3DShot1.
3) Assembly Selection and Refinement: 3DShotMQ reports the first assembled model, after refinement, as the final prediction.

3DShot2:
3DShot2 is a new autonomous server that generates hybrid models using models generated by Inub and our local implementations of HHpred and Sp3. Variable number of models are selected based on the target difficulty determined using the %sequence identity of the top predicted models. 3DShot2 generates hybrid models using an improved version of the shotgun algorithm and by doing assembly at the sub-structural unit level instead of the residue level followed by refinement.

AMU-Biology

Combined methods of template-based and template-free modeling

J.M. Kasprzak, T. Puton, M. Musielak, K. Milanowska, K. Majorek, M. Domagalski, E. Kubiaczyk, A. Wysoczanska, U. Baraniak, N. Szostak, M. Magnus, J.M. Bujnicki and A. Czerwoniec*
Bioinformatics Laboratory, Institute of Molecular Biology and Biotechnology, Adam Mickiewicz University, Umultowska 89, PL-61-614 Poznan, Poland
anna.czerwoniec@amu.edu.pl

In the eighth Critical Assessment of techniques for protein Structure Prediction (CASP8), the AMU-Biology group used the combination of the 'FRankenstein's Monster' approach for template-based modeling (Kosinski, 2003) with the REFINER tool (Boniecki, 2003) and the ROSETTA method for de

novo modeling (Simons, 1997) to predict the tertiary structure of full-length targets of all categories.

In the first step we identified structural homologs and generated target-template alignments using a number of fold-recognition methods available via the GeneSilico MetaServer (Kurowski and Bujnicki, 2003). MODELLER (Fiser and Sali, 2003) was used to convert target-template alignments into preliminary models. At this stage we also used external information: secondary structure predictions, conservation of fragments and putative catalytic residues, and constraints on the placement of insertion and deletions in the loop regions as well as available literature information. The preliminary models were evaluated according to MetaMQAP (Pawlowski, 2008 in press) to enable discrimination of fragments that are likely to be erroneous and ranked with automated ProQ assessment (Wallner, 2005).

After superimposing the best models and merging their best-scored fragments, we constructed hybrid models and used them to introduce necessary modifications in the original target-template alignments. The first step of model refinement involved iterative model building, evaluation, and realignment. The second step of model refinement concerned short loop regions, for which the structural information available from the templates was insufficient. We used REFINER (Boniecki, 2003) – a "*de novo*" protein structure prediction method. For each fragment we generated hundreds of alternative models, which were then ranked by ProQ to select the best models. In a few cases we used also the server for short loop modeling (Michalsky, 2003).

For long regions (or entire proteins) with no corresponding structure among the templates identified by fold-recognition, we attempted *de novo* modeling using the ROSETTA algorithm. Hundreds to thousands of decoys were generated and clustered to identify the most representative low-energy conformations. Models were selected according to the average energy of clusters, size, density and visual evaluation of the full-atom structures. The final hybrid models were 'refined' by running MODELLER to optimize the bond lengths and angles.

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 modeling: merging the finest fragments of fold-recognition models and iterative model refinement aided by 3D structure evaluation. Proteins 53 Suppl 6:369-79.
2. Simons KT, 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(1):209-25.
3. Boniecki M, Rotkiewicz P, Skolnick J, Kolinski A. (2003) Protein fragment reconstruction using various modeling techniques. J Comput Aided Mol Des. 2003 Nov;17(11):725-38.
4. Kurowski, M.A., Bujnicki J.M. (2003) GeneSilico protein structure prediction meta-server. Nucleic Acids Res. 31(13):3305-7.
5. Fiser, A., Sali, A. (2003). Modeller: generation and refinement of homology-based protein structure models. Methods Enzymol. 374:461-91.
6. Pawlowski M, Gajda MJ, Matlak R, Bujnicki JM. (2008 in press). MetaMQAP: a meta-server for the quality assessment of protein models. BMC Bioinformatics
7. Wallner B., Elofsson A (2005). Identification of correct regions in protein models using structural, alignment and consensus information. Protein Sci., 15(4):900-913
8. Michalsky E., Goede A., Preissner R. (2003) Loops In Proteins (LIP) – a comprehensive loop database for homology modeling.

BAKER-GINZU

Ginzu homolog identification and domain parsing in CASP8

D. Chivian[1], D. E. Kim[2], J. Thompson[2] and D. Baker[2]
[1] - *Lawrence Berkeley National Laboratory, Berkeley, CA,*
[2] - *University of Washington, Seattle, WA*
DCChivian@lbl.gov

Protein chains often contain more than one domain. In order to predict the domain organization of a protein, we have developed the Ginzu1-2 homolog identification and domain parsing method. The method is available to the public as part of the Robetta server1;3-4 (http://robetta.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/PSI-BLAST5 search against the PDB sequence database detects such relatives. All PDB ids that are detected at this stage are stored. 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 HHSEARCH6 in this step for the parsing of the CASP8 targets. Again, as with the PSI-BLAST detections, the associated PDB and region of the target chain covered is retained.

Any remaining long regions of the query that do not have structural homologs may require further division into domains. One may search unassigned regions against Pfam7. 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 likelihood of possessing 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 PSIPRED8, 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 method9 that is applied to the target's final Robetta model, as well as PSI-BLAST detectable structural homologs, to complete the domain parsing. Alternate domain predictions based on the model from the default K*Sync alignment to the parent are also returned, as are MSA-based predictions for weak confidence HHSEARCH detected regions.

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, 524-533.
2. Kim D.E., Chivian D., Malmstrom L., & Baker D. (2005) Automated prediction of domain boundaries in CASP6 targets using Ginzu and RosettaDOM. Proteins 61, 193-200.
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. Chivian D., Kim D.E., Malmstrom L., Schonbrun J., Rohl C.A., & Baker D. (2005) Prediction of CASP6 structures using automated Robetta protocols. Proteins 61, 157-166.
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. Soding J. (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics 21, 951-960.
7. 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.
8. Jones D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292, 195-202.
9. Taylor W.R. (1999) Protein structural domain identification. Protein Eng 12, 203-216.

Hybrid domain parsing with Ginzu and RosettaDOM

D. Chivian[1], D. E. Kim[2] and D. Baker[2]
[1] - *Lawrence Berkeley National Laboratory, Berkeley, CA,*
[2] - *University of Washington, Seattle, WA*
DCChivian@lbl.gov

Protein chains often contain more than one domain. In order to predict the domain organization of a protein, we have combined the Ginzu1-2 and RosettaDOM2 domain parsing methods into a hybrid predictor (see accompanying abstracts for BAKER-GINZU and BAKER-ROSETTADOM in this volume).

Ginzu attempts to determine the locations of putative domains in the query sequence and the identification of any likely homologs with experimentally characterized structures with PSI-BLAST3 and HHSEARCH4. This search for homologous structures is followed by parsing any remaining regions by screening Pfam5, and then by application of a boundary preference function. The boundary preference function is derived from a PSI-BLAST MSA (from the "UniRef90" sequence database6) via a heuristic that considers clusters of sequences in the PSI-BLAST MSA, the least occupied positions in the MSA, strongly predicted loop regions by PSIPRED7, 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. Regions with structural homologs are further parsed using a consensus variant of Taylor's structure-based domain parsing method8.

RosettaDOM generates 400 decoy structures with Rosetta's de novo fragment-assembly approach for the full length of the target and structurally parses each of those decoys using Taylor's structure-based domain parsing method. Increased frequency of boundaries within a sliding window (smoothed in the same fashion as SnapDRAGON9) is used to assign domain boundaries (over a Z-score of 2.5). Although Rosetta is unlikely to produce accurate atomic-resolution models, it may accurately produce course structural features such as domains.

Both Ginzu and RosettaDOM often do not arrive at a strongly predicted boundary separately, but instead may suggest several candidate boundaries with a confidence below the threshold of each method. In such circumstances, agreement between the two methods increases the confidence of a boundary within that window. The BAKER-DP_HYBRID method takes advantage of the agreement between the sequence-based and structure based domain prediction methods by combining the boundary confidence functions from the two methods (only in regions without a strongly detected PDB homolog by Ginzu). It reports boundaries only when the combined function is above the threshold, which may be achieved with a strong prediction by either method or when weaker predictions by each method are in agreement. Regions with PDB homologs found by Ginzu are structurally parsed with Taylor's method (based on the model) in the same fashion as Ginzu. The BAKER-DP_HYBRID method makes use of two largely independent domain prediction methods, one based on sequence homology and the other based on de novo structure predictions.

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, 524-533.
2. Kim D.E., Chivian D., Malmstrom L., & Baker D. (2005) Automated prediction of domain boundaries in CASP6 targets using Ginzu and RosettaDOM. Proteins 61, 193-200.
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. Soding J. (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics 21, 951-960.
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. Nucleic Acids Res 30, 276-280.
6. Suzek B.E., Huang H., McGarvey P., Mazumder R., & Wu C.H. (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 23, 1282-1288.
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.
9. George R.A., Heringa J. (2003) SnapDRAGON: a method to delineate protein structural domains from sequence data. J Mol Biol. 316, 839-851.

Robetta De Novo and Homology Modeling in CASP8

D. E. Kim[1], D. Chivian[2] , J. Thompson[1], R. Vernon[1] and D. Baker[1]
[1] - *University of Washington, Seattle, WA,*
[2] - *Lawrence Berkeley National Laboratory, Berkeley, CA*
dekim@u.washington.edu

The Robetta server[1-3] (http://robetta.org) combines the Rosetta homology modeling[4] and *de novo*[5-7] tertiary structure prediction protocols with the Ginzu[1,8] homolog identification and domain parsing protocol to provide predictions for the full length of each target. The main modifications to the Robetta homology modeling protocol for CASP8 include generating more conservative alignment ensembles for close homologs, and generating more compact loops and using a more stringent chain-break filter in loop modeling. As in CASP7, our model ensembles are parametrically generated for up to 5 parents by the K*Sync[4] alignment method for the template regions and with Rosetta loop modeling[7,9] for unaligned regions. The main modifications to the Robetta *de novo* protocol include the addition of a full-atom refinement step at the end of each standard Rosetta fragment-replacement trial, and a significant increase in the number of independent trajectories sampled through the use of the distributed computing project Rosetta@home (http://boinc.bakerlab.org/rosetta). Blind benchmarking of servers is crucial as it allows us to measure the abilities of automated prediction, which is vital for the purpose of large-scale prediction efforts.

Robetta homology modeling protocol
Robetta uses up to 5 of the highest confidence detections from BLAST/PSI-BLAST[8] or HHSEARCH[10] 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 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[4] 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 threading the sequence of the query onto the backbone of the parent based on the alignment. Unaligned regions are modeled using the standard Rosetta loop modeling protocol that involves cyclic coordinate descent and optimized to fit the aligned template structure[7,9]. The template region is kept fixed, and models are selected from the ensemble using a combination of the Rosetta energy function with a consensus score derived from the alignment ensemble[4].

Robetta *de novo* protocol
As in CASP7, Robetta *de novo* modeling generates 4000 query decoys and 2000 decoys each for up to 2 homologous sequences (filtered down to 2000, 1000, 1000 to ameliorate known pathologies such as low contact-order structures) using the Rosetta fragment-assembly methodology[5]. For CASP8, in an effort to obtain high-resolution models, we also generated up to 300,000 query decoys using the standard fragment replacement strategy followed by full-atom refinement using the Rosetta full-atom energy function[6,7]. The CPU cycles necessary for large scale conformational sampling and full-atom refinement was provided by the distributed computing project Rosetta@home. The lowest scoring 4000 decoys based on their full-atom energies were structurally clustered with the standard query and homolog decoy sets, and the lowest scoring full-atom decoy from each of the 5 best clusters by population were returned as the final predictions. The final predictions were ranked entirely based on the Rosetta all-atom energies.

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, 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. Chivian D., Kim D.E., Malmstrom L., Schonbrun J., Rohl C.A., & Baker D. (2005) Prediction of CASP6 structures using automated Robetta protocols. Proteins 61, 157-166.
4. Chivian D. & Baker D. (2006) Homology modeling using parametric alignment ensemble generation with consensus and energy-based model selection. Nucleic Acids Res. Sep 13 [Epub].
5. 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.
6. Bradley P., Misura K. M., Baker D. (2005). Toward high-resolution de novo structure prediction for small proteins Science 309, 1868-1871.

7. Das R., Qian B., Raman S., Vernon R., Thompson J., Bradley P., Khare S., Tyka M. D., Bhat D., Chivian D., Kim D. E., Sheffler W. H., Malmstrom L., Wollacott A. M., Wang C., Andre I., & Baker D. (2007) Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. Proteins 69, 118-128.
8. Kim D.E., Chivian D., Malmstrom L., & Baker D. (2005) Automated prediction of domain boundaries in CASP6 targets using Ginzu and RosettaDOM. Proteins 61, 193-200.
9. Canutescu, A.A. & Dunbrack, R.L., Jr. (2003). Cyclic coordinate descent: A robotics algorithm for protein loop closure. Protein Sci 12, 963-72.
10. Soding J. (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics 21, 951-960.

BAKER-ROSETTADOM

The RosettaDOM Domain Parsing Protocol

D. E. Kim[1], D. Chivian[2]and D. Baker[1]
[1] - *University of Washington, Seattle, WA,*
[2] - *Lawrence Berkeley National Laboratory, Berkeley, CA*
dekim@u.washington.edu

Here, we describe a protocol to identify protein domain boundaries using a sequence homology based procedure called Ginzu[1-2], and a *de novo* 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 or PSI-BLAST[6], RosettaDOM simply returns the domain boundary predictions provided by Ginzu. For query sequences lacking such homology, a *de novo* domain prediction method similar to SnapDRAGON[7] is used. The *de novo* 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[8]. 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 atomic-resolution models, it may accurately produce coarse structural features such as domains.

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. Km D.E., Chivian D., Malmstrom L., & Baker D. (2005) Automated prediction of domain boundaries in CASP6 targets using Ginzu and RosettaDOM. Proteins 61, 193-200.
3. Badley 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. George R.A. & Heringa J. (2002) SnapDRAGON: a method to delineate protein structural domains from sequence data. J. Mol. Biol. 316, 839-851.
8. Taylor W.R. (1999) Protein structural domain identification. Protein Eng. 12, 203-216.

Semi-automated tertiary structure prediction and ligand binding site prediction using in-house server based on fold recognition, fragment assembly, and quality assessment

S. Nakamura[1], M. Morita[1] and K. Shimizu[1]
1 - Department of Biotechnology, The University of Tokyo
shugo@bi.a.u-tokyo.ac.jp

As in CASP7, we used our in-house protein structure prediction server named "ENABLE" for tertiary structure prediction. The following is the overview of the prediction procedure: 1) Templates for a target were first searched using PDB-BLAST, FUGUE[1], and SP3[2]. About 20-30 3D models were generated based on various combinations of templates and alignments using MODELLER[3] and SCWRL[4]. 2) Qualities of the models were then assessed using our developed QA predictor. 3) If refinement of partial structure or full de novo prediction is needed, our developed de novo prediction tool named "IDDD/ABLE[5]" was executed. Target function including burial of hydrophobic residues, contacts between residues, average distance between hydrophobic residues, hydrogen bonds between mainchains, and exclusive volume to avoid overlap of residues was minimized by simulated annealing with 40000 steps. About 5000-20000 models were generated using IDDD/ABLE according to the length of a target and the available computational resources. 4) Apply clustering to generated models and five models were picked up that have best QA score and were within 5 % from the center of the largest clusters. Our QA predictor is based on support vector regression (SVR) and predict GDT-TS and TM-score[6] of a model in global mode and can predict S-score for each residue of a model in local mode using scores of Verify3D[7], ProSa[8], ProQ[9], secondary structure matching using PSIPRED[10], and IDDD/ABLE potentials.

For ligand binding site prediction, server models were first collected from CASP8 web site and qualities of the models were assessed using our QA predictor, and five model structures were selected from top of the QA list semi-automatically. Then ligand-binding sites for each model were predicted by the structure-based approach[11], followed by the sequence-based approach. In the structure-based approach, the protein surface was coated with multiple layers of probes to calculate the van der Waals interaction energies between these probes and the protein. Energetically favorable probes were then clustered and the resulting clusters were ranked based on their total interaction energies. In the following sequence-based approach, the first three clusters, which were generated by the structure-based approach, were re-ranked according to average conservation scores of all residues within 4 Å from any probe in each cluster. The first ranked cluster was regarded as the predicted site for a model. Finally, predicted sites (one site per model, five sites in total) were submitted in the order of the same score as was used in sequence-based approach.

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. 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.
3. Sali,A. & Blundell,T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234, 779-815.
4. Canutescu,A.A., Shelenkov,A.A. & Dunbrack R.L. (2003). A graph theory algorithm for protein side-chain prediction. Protein Sci. 12, 2001-2014.
5. Ishida,T., Nishimura,T., Nozaki,M., Inoue,T., Terada,T., Nakamura,S. & Shimizu,K. (2003). Development of an ab initio protein structure prediction system ABLE. Proc. 14th Int'l Conf. Genome Inform. (GIW 2003) 14, 228-237.
6. Zhang, Y. & Skolnick, J. (2004) Scoring function for automated assessment of protein structure template quality. Proteins 57, 702-710.
7. Luthy,R., Bowie,J.U. & Eisenberg,D. (1992). Assessment of protein models with three-dimensional profiles. Nature 356, 83-85.
8. Sippl,M.J. (1993). Recognition of Errors in Three-Dimensional Structures of Proteins. Proteins 17, 355-362.
9. Wallner,B. & Elofsson,A. (2003). Can correct protein models be identified? Protein Sci. 12, 1073-1086.
10. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292, 195-202.
11. Morita, M., Nakamura, S. & Shimizu, K. (2008) Highly accurate method for ligand-binding site prediction in unbound state (apo) protein structures. Proteins 73, 468-479.

PID-SVM: Prediction of Intrinsic Disordered Regions Using Multiple Sequence-Derived Inputs and Customized Models

K. Chen[1], H. Zhang[1,2] and L. Kurgan[1]
[1] - *Electrical and Computer Engineering, University of Alberta, Canada*
[2] - *College of Mathematical Science and LPMC, Nankai University, PRC*
lkurgan@ece.ualberta.ca

We present a method, PID-SVM, for the sequence-based prediction of intrinsic disordered regions. The overall architecture of PID-SVM, which incorporates two steps, is similar to other existing disorder prediction methods. First, each predicted residue is represented by a fixed-length numerical feature vector. The vector is extracted using a 15-residue wide window which is centered on the predicted residue. Second, the vector is inputted into a support vector machine (SVM) classifier that outputs the prediction, i.e., ordered or disordered residue, together with a numerical score that is normalized to [0, 1] interval. The novelty of the PID-SVM method stems from the following four characteristics.

1. The features that implement the input to the SMV integrate three sequence-derived sources, the PSI-BLAST[1] profile (PSSM), the secondary structure predicted with PSI-PRED[2], which is represented by the probabilities of assuming the helix, strand, and coil conformations, and the output of IUPred[3], which represent the probability that a given residue is disordered.

2. The method is custom-designed to differentiate between predictions for residues at the sequence termini (the first and the last 20 amino acids in the sequence) and for the remaining internal (with respect to the sequence) residues. This is motivated by an observation that majority of the disordered regions are located in the vicinity of the sequence termini, which calls for a different (biased towards prediction of the disordered residues) prediction model when compared with the model for the internal residues (biased towards prediction of ordered residues). Although we encode all residues using the same set of features, the training of the model is used to accommodate for the bias. To this end, each feature f is duplicated into two features, f_t (termini) and f_i (internal). For the residues at the termini, we set $f_t = f$ and $f_i = 0$, while for the internal residues $f_t = 0$ and $f_i = f$.

3. PID-SVM applies a cost matrix to accommodate for the unbalanced ratio between disordered and ordered residues, i.e., lower weigh values are associated with the majority (ordered) residues. More specifically, the disordered regions occupy only about 5%-6% of the protein sequences[4]. Classifiers that treat the majority (larger) and minority (smaller) classes the same way develop a bias toward the majority class resulting in low recall for the minority class and a low ROC value. Usage of the cost matrix allows balancing the predictions between the ordered and the disordered residues.

4. PID-SVM applies two feature selection methods, the information gain based method and the Chi-squared method, to reduce the dimensionality (allowing for more efficient training of the classification model) and to select features that contribute to the improved classification performance (potentially improving the prediction quality by removing irrelevant features). The selection methods rank the features based on two different merit measures, which were shown to be complementary[5]. The features were sorted based on their average rank and the top 100 features were selected as the input to the SVM classifier.

The PID-SVM does not utilize structural templates, which allows for consistent performance for both homologous and low-homology targets. The method was trained using sequences extracted from PDB. The pairwise sequence identity in the training set is below 25% to assure that the generated classification model can generalize to sequence with low identity.

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. Dosztányi,Z., Csizmok,V., Tompa,P. & Simon,I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21**, 3433-4.
4. Bordoli,L., Kiefer,F. & Schwede,T. (2007). Assessment of disorder predictions in CASP7. *Proteins* **69** Suppl 8, 129-36.
5. Chen,K., Jiang,Y., Du,L. & Kurgan,L. (2008). Prediction of integral membrane protein type by collocated hydrophobic amino acid pairs. *J. Comp. Chem.* DOI: 10.1002/jcc.21053.

CADCMLAB

Combining Spectral Based Sequence Comparison Methods with Orthodox Sequence Alignment Techniques for Protein Fold Recognition and 3-D Structure Prediction

C.A. Del Carpio[1], I. Mohamed[1], E. Ichiishi[1,2], H. Tsuboi[1], A. Endou[1], H. Takaba[1] and A. Miyamoto[1]

[1] - *Graduate School of Engineering, Tohoku University. Sendai Japan.*

, [2] - *Japan Advanced Institute of Science and Technology, Kanazawa, Japan.*

carlos@aki.che.tohoku.ac.jp

For CASP8, we have combined our original system for protein 3D structure prediction PIPS[1,2] with orthodox sequence alignment techniques. The underlying concept in the spectral analysis method embedded in PIPS is 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[1,2] by means of which the cepstrum (measure of the periodic wiggliness of a frequency response) is computed so as to lead to a spectral envelope that depicts the subtle periodicity in physicochemical characteristics of the amino acid sequences. The system extracts a diverse set of proteins from PDB when the methodology is applied to a target sequence in order to search similar folding patterns. Extracted structures rank from scant similarity in terms of amino acid composition to high similarity ones. Then a more specific sequence alignment like FASTA (http://www.ebi.ac.uk/Tools/fasta33) or BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi) can be applied to the reduced set of structures obtained by our spectral oriented methodology. This combined method has shown a high degree of effectiveness to select optimal templates for a determined target, both in terms of processing times as well as quality of template. The threading algorithm is then pursued by an energy minimization process for the newly built structure. Table 1 shows a list of the targets in which the methodology has succeeded in recognized the closest folding pattern for the targets in CASP8.

1. Del Carpio, C.A. and Yoshimor, A. (2002) International University Line Publishers (IUL), 171-200.
2. Del Carpio, C.A. and Carbajal, J.C. (2002) Genome Informatics 13, 163-172..

Table 1. Comparison of our group's results for some CASP8 targets

N	Target	Length	Fitted length	RMSD
1	T0389_1	153	119	1.667
2	T0393_2	263	147	2.511
3	T0401_1	143	108	1.952
4	T0407_1	363	177	2.064
5	T0411_1	139	101	1.621
6	T0415_1	109	95	1.979
7	T0417_2	189	125	2.332
8	T0423_1	156	142	1.098
9	T0425_1	181	164	1.830
10	T0427_1	422	328	2.424
11	T0431_1	491	427	1.554
12	T0434_5	205	106	1.535
13	T0437_1	99	68	4.441
14	T0440_1	275	231	1.576
15	T0446_1	124	96	2.254
16	T0449_1	307	225	1.608
17	T0451_1	133	99	2.443
18	T0454_1	203	147	2.401
19	T0469_1	65	31	2.235
20	T0480_3	55	26	1.783
21	T0492_1	73	46	2.020

TESE: Generating specific protein structure test set ensembles

F. Sirocco[1] and S.C.E. Tosatto[1]
[1] - *Department of Biology, University of Padova*
silvio.tosatto@unipd.it

Creating representative ensembles of sufficiently diverse proteins is a recurring problem in bioinformatics. Any novel method has to be trained and benchmarked on a test set of protein sequences and/or structures ensuring wide coverage of the protein universe and solid statistical evaluation. At least three different use cases can be envisaged: (i) The benchmarking of novel sequence alignment protocols and statistical potentials. (ii) The generation of test sets for specialized protein classes, e.g. transmembrane proteins. (iii) Extending datasets from previous publications with new structures to enhance statistical significance, e.g. for novel repeat proteins. Given the exponential growth in available information, it is increasingly necessary to generate representative test sets large enough to allow solid statistical evaluation of the results.

One limitation of currently available services is the lack of an underlying structural classification throughout the selection process. This becomes increasingly important in the low sequence similarity range, where it is desirable to eliminate homology, and limits the usefulness of current methods in fold recognition for instance. On the other hand, the structural classification schemes, e.g. CATH[1], are readily used for the selection of similar structures in absence of sequence similarity. However, only the full classifications are distributed and it is the developer's responsibility to extract meaningful subsets in a similar way to the previously mentioned services. This process can become rather cumbersome in practice, e.g. when selecting structures with short tandem repeats or representatives of the Rossman fold. A lack of standardization, and the relevance of many technical details in the selection process, frequently also complicates the unbiased assessment of novel methods to avoid "cherry-picking" of the data. For these reasons, we have developed TESE[2], a novel server for the automatic generation of large benchmark sets both on the sequence and on the structure level.

TESE is a method to derive meaningful ad hoc test sets from proteins of known structure. The CATH structural classification is used to control sequence/structural redundancy at various levels, e.g. <35% pairwise sequence identity corresponds to the "S" level. Queries may be started in three different ways. Keywords or a small sample of PDB files can be used to seed the TESE search for specific proteins, e.g. for alpha-helical repeats or oxidoreductases, or to extend previously published datasets. Alternatively, the user may specify search parameters related to the desired CATH similarity level, e.g. topology, the experimental method and quality, e.g. maximum X-ray resolution, or protein size, e.g. minimum length, to initiate the search. It is possible to select all structures or a randomly chosen subset of any size. For sets of less than 600 proteins, a clickable list of protein structures and their CATH classification is produced. New proteins may be selected by directly choosing a different protein subset or by adding additional search parameters. When satisfied, the user may save the protein list as a compressed archive containing the relevant FASTA formatted sequences, PDB files and a HTML index of the selected proteins. The test set may be automatically split to create subsets for cross-validation. Large datasets of more than 600 proteins are treated in a non-interactive way to limit bandwidth usage. Some widely used test sets are available as precompiled archives. An online help is provided to guide the user through the process. A more extensive server description and examples are available from the web site at URL: http://protein.bio.unipd.it/tese/.

1. Sirocco, F., Tosatto, S.C. (2008) TESE: Generating specific protein structure test set ensembles. Bioinformatics. 2008 Sep 16. [Epub ahead of print]
2. Pearl, F.M., Bennett, C.F., Bray, J.E., Harrison, A.P., Martin, N., Shepherd, A., Sillitoe, I., Thornton, J. and Orengo, C.A. (2003) The CATH database: an extended protein family resource for structural and functional genomics, Nucleic Acids Res, 31, 452-455.

CaspIta

Prediction of intrinsically disordered regions with ASPIDES

G. Zamperin[1], A. Sperduti[2] and S.C.E. Tosatto[1]
[1] - *Department of Biology,* [2] - *Department of Pure and Applied Mathematics, University of Padova*
silvio.tosatto@unipd.it

ASPIDES (Atchley and Svm Predict Intrinsically Disorder ElementS) is a software that predicts disorder in proteins sequence. It uses the five physico-chemical property scales of Atchley et al.[1] in order to translate the protein sequence into numbers. The prediction is made with a support vector machine trained in analogy to SPRITZ[2].

From the input sequence, ASPIDES creates a multiple sequence alignment with PSI-BLAST[3], secondary structure prediction with Porter[4] and solvent accessibility prediction with ACCpro[5]. These three are melded together to create the features: secondary structure and solvent accessibility predictions are considered as they are, while the multiple sequence alignment is processed using the five physico-chemical property scales of Atchley.

Described in this way, each residue is given in input to the support vector machine. The prediction is made using a Gaussian kernel. The output is filtered using the PDB[6] in order to reduce the output probability in case of many similar sequences in PDB, corresponding to a higher probability for the residue to be ordered rather than disordered. The resulting values are the output of the predictor ASPIDES.

1. Atchley WR, Zhao J, Fernandes AD, Drüke T. (2005) Solving the protein sequence metric problem. Proc Natl Acad Sci USA, 102(18):6395-400.
2. Vullo A, Bortolami O, Pollastri G and Tosatto SC (2006) Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines. Nucleic Acids Res, 34:W164-168.
3. 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.
4. Pollastri, G. and McLysaght, A. (2005) Porter: a new, accurate server for protein secondary structure prediction, Bioinformatics, 21, 1719-1720.
5. Pollastri, G., Baldi, P., Fariselli, P. and Casadio, R. (2002) Prediction of coordination number and relative solvent accessibility in proteins, Proteins, 47, 142-153.
6. Berman, H.M., Battistuz, T., Bhat, T.N., Bluhm, W.F., Bourne, P.E., Burkhardt, K., Feng, Z., Gilliland, G.L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J.D. and Zardecki, C. (2002) The Protein Data Bank, Acta Crystallogr D Biol Crystallogr, 58, 899-907.

CBRC-DP_DR

Protein disordered region and domain prediction by using POODLE-I and domain linker prediction methods

T. Noguchi, S. Yamada, K. Shimizu and S. Hirose
Computational Biology Research Center
National Institute of Advanced Industrial Science and Technology, Japan
noguchi-tamotsu@aist.go.jp

We predicted disordered regions and domains in proteins by using POODLE-I, which is the disordered region prediction server (see the abstract for CBRC_POODLE in this volume), and two methods for domain linker prediction. In the previous CASP experiments, our team succeeded in improvement of prediction accuracy for disordered regions in protein by our original predictors (i.e. POODLE series: POODLE-S[1], L[2] and W[3]) in combination with prediction results of secondary structure, accessible surface area and similar structures. POODLE-I includes several predictors: POODLE series as disordered region prediction, PSIPRED[4], jpred[5] and sable[6] as secondary structure prediction, jnet and sable as accessible surface area prediction, genThreader[7] and HHsearch[8] as fold recognition, and coils[9] as coiled coil region prediction. DomCut[10] and our original method based on position-specific scoring matrix are used for the domain linker prediction. Our method was developed by *M.Takizawa,* who was a member of our group in

CASP7. This method was assessed using a dataset of 106 multi-protein domains defined in SCOP. The performance of our method for predicting domain linker, which exhibited sensitivity of 39.3% and specificity of 64.7%, was higher when compared with several other methods (i.e. DLP[11], DomCut and Armadillo[12]).

We respectively predicted disordered region and domain linker by using our original method, POODLE series and our domain linker prediction method, and then the prediction results were carefully inspected with reference to the template structure and/or the predicted secondary structure obtained from other programs, and the final disordered regions and domains were determined. In case of detecting a high scored template structure by fold recognition, the information had priority for modifying predicted the disordered regions and the domains. And in case that target sequence were not covered by the template structure or that a template was not found, we give priority to the disordered region prediction of POODLE series and domain prediction of our method based on position-specific scoring matrix respectively.

1. Shimizu K., Hirose S. & Noguchi T. (2007). POODLE-S: web application for predicting protein disorder by using physicochemical features and reduced amino acid set of a protein-specific scoring matrix. Bioinformatics, 23(17), 2337-2338.
2. Hirose S., Shimizu K., Kanai S., Kuroda Y. & Noguchi T. (2007). POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions. Bioinformatics, 23(17), 2046-2053.
3. Shimizu K., Muraoka Y., Hirose S., Tomii K. & Noguchi T. (2007). Predicting mostly disordered proteins by using structure-unknown protein data. BMC Bioinformatics, 8, 78.
4. McGuffin J., Bryson K. & Jones DT. (2000). The PSIPERD protein structure prediction server. Bioinformatics, 16(4), 404-405.
5. Cuff JA. & Barton GJ. (2000). Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins, 40(3), 502-511.
6. Adamczak R., Porollo A. & Meller J. (2005). Combining prediction of secondary structure and solvent accessibility in proteins. Proteins, 59(3), 467-475.
7. Jones DT. (1999). GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J. Mol. Biol., 287(4), 797-815.
8. Söding J. (2005). Protein homology detection by HMM-HMM comparison. Bioinformatics, 21(7), 951-960.
9. Lupas A., Van Dyke M. & Stock J. (1991). Predicting coiled coils from protein sequence. Science, 252(5009), 1162-1164.
10. Suyama,M. & Ohara,O. (2003). DomCut: prediction of inter-domain linker regions in amino acid sequences. Bioinformatics, 19, 673-674.
11. 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.
12. Dumontier M., Yao R., Feldman HJ. & Hogue CW. (2005). Armadillo: domain boundary prediction by amino acid composition, J. Mol. Biol., 350, 1061-1073.

CBRC_POODLE

Disordered region prediction by integrating POODLE series

S. Hirose[1], K. Shimizu[1], N. Inoue[2], S. Kanai[2] and T. Noguchi[1]
[1] – Computational Biology Research Center (CBRC),
National Institute of Advanced Industrial Science and Technology, Japan
[2] – PharmaDesign, Inc., Japan
poodle@cbrc.jp

POODLE-I (“I” stands for Integration) is the disordered region prediction server based on integration of POODLE series. POODLE series consists of three programs that they target different disordered region according to their length[1-3]. In the previous CASP experiments, our team succeeded in improvement of prediction accuracy by adding prediction results of secondary structure, accessible surface area and similar structures. In this time, POODLE-I automated this technique.

POODLE-I includes several predictors: POODLE-S (two disordered definition are adopted; missing region and high B-factor region), L and W as disordered region prediction, PSIPRED[4], jpred[5] and sable[6] as secondary structure prediction, jnet and sable as accessible surface area prediction, genThreader[7] and HHsearch[8] as fold recognition, and coils[9] as coiled coil region prediction.
Our prediction method in POODLE-I consists of three steps, which are prediction, integration, and validation/modification step. The first step is to execute all prediction programs for a query sequence, and to align prediction results. The second step is to integrate prediction results of POODLE series. All

disordered regions predicted by POODLE series were picked up, and they were mapped on the prediction result of POODLE-S. But, disordered region with less than 30 aa length predicted by POODLE-L was ignored. The third step is to verify and modify predicted disordered region. Three kinds of rules were applied according to the length of disordered region. (i) If POODLE-W predicts that a query is unfolded protein, POODLE-I employed this result. (ii) If POODLE-L predicted long disordered region in a query, both terminal of disordered region were modified that they did not contain alpha-helix and beta-sheet based on information of predicted secondary structure. Exceptionally, if long disordered region was coincident with predicted coiled coil region, it was converted into ordered region. (iii) If short disordered region was predicted by POODLE-S, it was judged whether it was reliable prediction by using information of protein structure. If similar structures are able to be detected, it is confirmed whether there are any missing region or insertion in the corresponding region that was predicted disordered. In the case of no information of similar structure, we used the information of predicted secondary structure and accessible surface area. If a disordered region matched secondary structure or buried region, it was converted into an ordered region. Finally, if missing region(s) was detected in the predicted ordered region, an ordered region was converted into a disordered region.

All information about the POODLE series is provided at http://mbs.cbrc.jp/poodle.

1. Shimizu K., Hirose S. & Noguchi T. (2007). POODLE-S: web application for predicting protein disorder by using physicochemical features and reduced amino acid set of a protein-specific scoring matrix. Bioinformatics, 23(17), 2337-2338.
2. Hirose S., Shimizu K., Kanai S., Kuroda Y. & Noguchi T. (2007). POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions. Bioinformatics, 23(17), 2046-2053.
3. Shimizu K., Muraoka Y., Hirose S., Tomii K. & Noguchi T. (2007). Predicting mostly disordered proteins by using structure-unknown protein data. BMC Bioinformatics, 8, 78.
4. McGuffin J., Bryson K. & Jones DT. (2000). The PSIPERD protein structure prediction server. Bioinformatics, 16(4), 404-405.
5. Cuff JA. & Barton GJ. (2000). Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins, 40(3), 502-511.
6. Adamczak R., Porollo A. & Meller J. (2005). Combining prediction of secondary structure and solvent accessibility in proteins. Proteins, 59(3), 467-475.
7. Jones DT. (1999). GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J. Mol. Biol., 287(4), 797-815.
8. Söding J. (2005). Protein homology detection by HMM-HMM comparison. Bioinformatics, 21(7), 951-960.
9. Lupas A., Van Dyke M. & Stock J. (1991). Predicting coiled coils from protein sequence. Science, 252(5009), 1162-1164.

Chicken_George

Template-based modeling and free-modeling by fragment assembly with SimFold energy function

K. Sawada, S. Minami, M. Yamaura, S. Sawada and G. Chikenji

Department of Computational Science and Engineering,
Graduate School of Engineering, Nagoya University
chikenji@tbp.cse.nagoya-u.ac.jp

In this round of CASP experiment, we used SimFold[1,] a protein structure prediction toolbox that we have been developing, and submitted all targets in the TS category. Regardless of the target difficulty, we performed fragment assembly[2] with SimFold energy function for all targets (even for easy targets). But, depending on the difficulty, fragment library construction procedures were different. Hereafter, we briefly describe the energy function and what we did in CASP8.

SimFold energy function

SimFold uses a reduced protein model representation that has explicit backbone atoms and a sphere at the center of mass of side-chain atoms. The energy function consists of several terms such as hydrophobic interaction, hydrogen bonding, and so on. Their functional forms are well based on physico-chemical consideration so that each energetic term can be interpreted as a physical force. The explicit expression of the energy function is described in ref [1].

The category classification

Before structure modeling, all targets were classified into easy/medium/hard categories, because we carried out different procedures for constructing fragment libraries depending on the target difficulty. First, the server predictions were downloaded from the CASP web site for each target. Second, we performed

structural clustering with TM-score[3] cut off 0.4. If the size of the largest cluster was larger than 40% of total number of server predictions, we assumed that the target was classified into the easy category. For the target in which we couldn't get sufficient cluster size, we used 3D-Jury system[4] which is clone software made by us. If the 3D-Jury system detected a template, we assumed that the target was classified into the medium category. Otherwise we classified the target into hard category.

Fragment library construction

(i) Easy targets:

Briefly, the concept of our method for easy targets is similar to that of the meta-predictor methods[5-7], in which 3D models are generated by hybridizing fragments of models obtained from several fold recognition servers, but the procedure of ours is much simpler than that of those methods. First, we performed structural clustering using TM-score. Cut off values were determined so that the size of the largest cluster is almost equal to 1/3 of the total number of server predictions. Second, all possible 10-residue fragments are excised from structures that are the members of the largest cluster. These fragment structures were used in the tertiary structure generation step as fragment libraries.

(ii) Medium targets:

Our attempt in this category was to refine the 3D-Jury top hit template by exhaustive search of conformational space that is near the template structure. To do so, fragment structures that are structurally similar to the template fragment were taken from PDB by the structure alignment program TM-align[8]. The criterion we used here was that fragments must be longer than 20-residue, and that RMSD to the corresponding region of the template must be smaller than 5 angstroms.

(iii) Hard targets:

For hard targets, we searched fragment structure candidates of 10-residue segment using the Pearson's correlation coefficient between the PSSMs of a query subsequence and the PSSMs of a target subsequence. Top 500 scoring structures are deposited in the fragment library for every overlapping 10-residue segment.

Tertiary structure generation

For tertiary structure generation, we performed fragment assembly combined with a simple simulated annealing algorithm for easy/medium targets, and with the replica exchange Monte Carlo for hard targets. Submitted models were basically selected from low energy structures by structure clustering, but if we couldn't obtain reasonable cluster size, we selected models by visual inspection. Finally, side-chain modeling was performed by using SCWRL version 3.0[9].

1. Fujitsuka, Y., Chikenji, G. & Takada, S. (2006). SimFold energy function for de novo protein structure prediction: Consensus with Rosetta. Proteins. 62, 381-398.
2. Simons, KT., Kooperberg, C., Huang, E., & Baker, D. (1997). In silico Protein Recombination: Enhancing Template and Sequence Alignment Selection for Comparative Protein Modeling. J. Mol. Biol. 268, 209-225.
3. Zhang, Y. & Skolnick, J. (2004). Scoring Function for Automated Assessment of Protein Structure Template Quality. Proteins. 57, 702-110.
4. Ginalski, K., Elofsson, A., Fischer, D., & Rychlewski, L. (2003). 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics. 19, 1015-1018
5. Fischer, D (2003) 3D-SHOTGUN: A Novel, Cooperative, Fold-Recognition Meta-Predictor. Proteins. 51, 434-441.
6. Kosinski J. et al. (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, 369-379.
7. Contreras-Moreira, B., Fitzjohn, P.W., & Bates, P.A. (2003). In silico Protein Recombination: Enhancing Template and Sequence Alignment Selection for Comparative Protein Modeling. J. Mol. Biol. 328, 593-608.
8. Zhang, Y. & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic. Acids. Res. 33, 2302-2309.
9. Canutescu, A.A, Shelenkov, A.A., Dunbrack, Jr. R.L. (2003). A graph theory algorithm for protein side-chain prediction. Protein Sci. 12, 2001-2014.

Development of Model Quality Assessment program using the secondary structure prediction and side-chain environment

G. Terashi[1], H. Sakai[1], K. Kanou[1], T. Hirata[1], M.Takeda-Shitaka[1], and H. Umeyama[1]
[1] - *School of Pharmacy, Kitasato University*
terashig@pharm.kitasato-u.ac.jp

In this work, we have developed MQ Assessment Programs CIRCLE [1] and participated in Quality Assessment (QA) category of CASP8. CIRCLE aims at identifying the near native models and incorrect models without using consensus methods.

CIRCLE considers two terms for the model quality: (1) model quality calculated from the side-chain environment of each residue (*SideChainScore* in equation(1)); and (2) similarity between the secondary structure propensities predicted for an amino acid sequence by PSI-PRED and the secondary structure of the three-dimensional model (*SSscore* in equation (1)). The side-chain environment for each residue is determined from the fraction of the molecular surface area of the side-chain covered by the polar atoms, the fraction of the side-chain area buried by any other atoms, and the secondary structure. According to the target difficulty, a total score is calculated as:

$$TotalScore = \begin{cases} \sum_{n}^{length} (0.35 \times SSscore + SideChainScore_{CM})_n & CM \\ \sum_{n}^{length} (0.75 \times SSscore + SideChainScore_{FRNF})_n & FRorNF \end{cases} \quad (1)$$

As shown in equation (1), the similarity score of the secondary structures (*SSscore*) is emphasized in difficult targets (FR: Fold Recognition, NF: New Fold) than easy targets (CM: Comparative Modeling).
In the QA category of CASP8, predictor groups provide quality estimates comprising scores between 0.0 and 1.0 for each protein structure model produced by server groups participating CASP8. Therefore, for each target, we convert estimated score of models into the values from 0.0 to 1.0 by scaling circle score of models which has minimum and maximum values.
The 103/128 (80%) native protein structures of CASP8 targets were published in CASP8 web site (Sep 2008). We calculated Pearson's correlation coefficient between converted CIRCLE score and the quality of models. We used the Global Distance Test Total Score (GDT_TS) as the quality of model compared to native.
The average of GDT_TS (x-axis) and correlation coefficient (y-axis) are shown in Fig.1. These results show that QA performance of CIRCLE depends on the quality of set of models which are evaluated (Table 1). The good correlation coefficients were obtained above 0.9 for the targets having the high average value of GDT_TS (above 50).

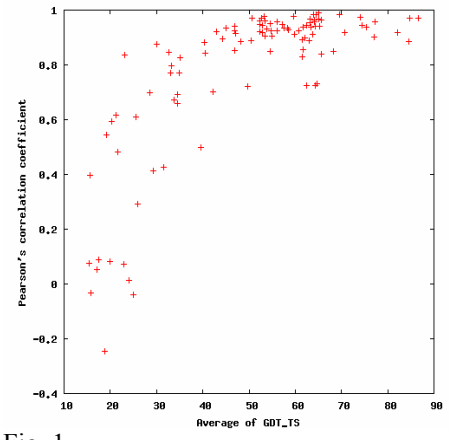


Fig. 1

Average of GDT_TS	Average of Pearson's correlation coefficient
0–25	0.24
25–50	0.75
50–75	0.92
75–100	0.93
ALL	0.78

Table 1

Additionally the best (T0423) and worst (T0460) examples of CIRCLE results are shown in Fig.2 and Fig.3. The x-axis and y-axis represents the circle score and GDT-TS of each model, respectively. In T0423 (Fig.2), CIRCLE score has high value of correlation coefficient (0.98), because high quality models (GDT_TS > 50) has high proportion of set of models. In contrast, in the case that no good models existed in the set of models (T0460 of Fig.3), CIECLE could not perform well (correlation coefficient = -0.24). These results indicate that the model accuracy of easy targets (GDT_TS>50) can be assessed quantitatively by our CIRCLE, and there is a room to improve especially in difficult targets, which does not include high quality

models. We are planning to add other kind of scoring function calculated from evolutional information such as a sequence alignment score and the consensus method.

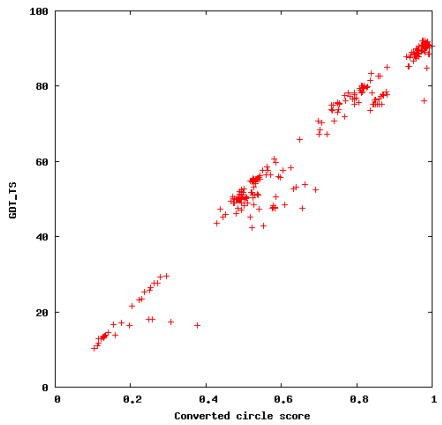


Fig. 2 T0423

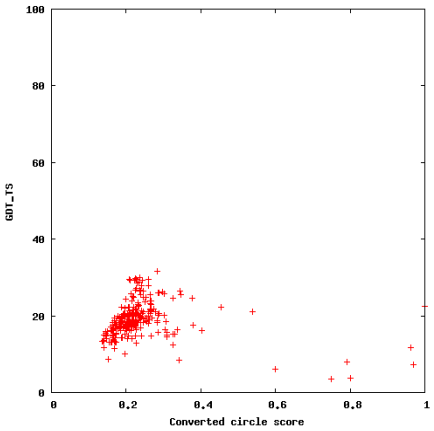


Fig. 3 T0460

1. Terashi G, Takeda-Shitaka M, Kanou K, Iwadate M, Takaya D, Hosoi A, Ohta K, Umeyama H.(2007) Fams-ace: a combined method to select the best model after remodeling all server models. Proteins. 69 Suppl 8:98-107.

COMA
COMA-M

Template-based modeling using COMA (*C*omparison *O*f *M*ultiple *A*lignments)

M. Margelevičius and Č. Venclovas
Institute of Biotechnology, Graičiūno 8, Vilnius, Lithuania
minmar@ibt.lt

COMA (*C*omparison *O*f *M*ultiple *A*lignments) is a novel profile comparison and search method (manuscript in preparation) that includes a number of new features. The major ones are: (1) filtering of non-informative profile regions with the modified two-level SEG algorithm, (2) composition-based statistics for profiles, (3) scoring schemes that could be used either to score independent pairs of profiles, or as a universal scoring system within context of a database, (4) variable (position-specific) gap penalties for the profile-profile alignment, (5) adaptive adjustment of the search and scoring parameters, (6) statistical significance estimation method derived specifically for profiles.

During CASP8 two versions of COMA-based automatic modeling servers were tested. Both servers build structural models using Modeller[1] from the alignments that are obtained by searching the target profile against the databases of profiles for structural templates. The differences between the two are in the processing of the profile search results. COMA models are always based on a single best template, while COMA-M is able to use multiple structural templates for model-building.

Target profiles are constructed from the alignments obtained using PSI-BLAST[2]. In every PSI-BLAST iteration the set of detected sequences is verified and the search parameters are appropriately adjusted. Several sequence databases are used for PSI-BLAST runs: nr80 (nr filtered at 80% of sequence identity), nr70, env_nr80 (junction of nr and env_nr filtered at 80%), and nr. Different databases can be selected during PSI-BLAST search depending on the results of the previous iteration. The search results also control whether sequence masking with SEG is turned on or off. Number of sequences for inclusion into the profile to be used in the next iteration is controlled at each iteration by analyzing the number of sequences found in the current iteration. The PSI-BLAST search is terminated either if it converges, a maximum number of iterations has been reached, or if at least one sequence included in the last sequence profile, is not found in the current iteration.

The obtained PSI-BLAST alignments are then used to make the target profiles using a program from the COMA toolkit. The profile for each target is searched in parallel against several databases of template profiles. These different databases include the same templates, but profiles are compiled using different sets of parameters. The results of these parallel searches are processed together.

Depending on the initial number of profiles found and their statistical significance, COMA either stops or continues the search in an iterative mode with the parameters adjusted on the fly. After the search results are obtained, the top 5 alignments with potential templates are extracted for each set of the results. All the possible combinations with different number of the top 5 alignments are then made. For each combination of the alignments, the templates are mutually aligned with DaliLite3. If a combination comprises more than 2 alignments, a multiple alignment is built using the DaliLite pairwise alignments. For each combination, a multiple alignment is constructed by taking COMA's alignment between the target and the template as a reference. Other templates are then added to the target-template alignment without changing it according to the DaliLite pairwise template comparisons. There are as many alignment variants as the number of templates in the combination.

After all variants of alignments are produced, serial modeling of the target with MODELLER is run using one or more structural templates (COMA always uses a single top template). Side chains of each model are optimized with SCWRL3[4].

The top models from the hundreds obtained are selected using the Prosa5 values. However, small variations in the model completeness may result in misleading Prosa Z-scores. Therefore, if two models overlap 90% or more, then the model with the better Prosa value for the overlapping region is selected.

1. Šali,A. & Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234, 779-815.
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. Holm,L. & Park,J. (2000) DaliLite workbench for protein structure comparison. Bioinformatics 16,566-567.
4. Canutescu,A.A., Shelenkov,A.A. & Dunbrack,R.L. (2003) A graph theory algorithm for protein side-chain prediction. Protein Science 12, 2001-2014.
5. Sippl,M.J. (1993) Recognition of Errors in Three-Dimensional Structures of Proteins. Proteins 17, 355-362.

CpHModels_193

CpHModels-3.0. Remote homology modeling using structure guided profile sequence alignments and double-sided baseline corrected scoring scheme

M. Nielsen, C. Lundegaard, O. Lund and T. N. Petersen

Center for Biological Sequence Analysis, Department of Systems biology, The Technical University of Denmark, Denmark

mniel@cbs.dtu.dk

Sequence profiles have a broad application in field of bioinformatics prediction algorithms dating back to the pioneering work by Rost and Sanders[1]. The field of protein structure prediction has largely benefited from this work, and most high performing algorithms for protein homology modeling use sequence profiles as their main vehicle[2-4]. Likewise has prediction of local protein structural features been demonstrated to improve when sequence profile are used to represent the protein sequences[5,6]. Here, we develop a scoring scheme for remote homology modeling building on these findings. Two protein sequences are aligned using local sequence alignment with an amino acids scoring matrix constructed combining sequence profiles, and local protein structural features like secondary structure and relative surface accessibility. For the query sequence where the structure is unknown, predicted local features are used. For the template PDB structure averages of predicted and DSSP assigned local features are used. Secondary structure predictions are performed using the artificial neural network approach described by Petersen et al [1], and relative surface exposure predicted using a doubled structure neural network approach as described by Petersen et al.[7]. Each element in the alignment function (profile, secondary structure, and relative surface exposure) where scored using a log-likelihood approach where the likelihood was estimated as $(\sum_a p_a^i \cdot p_a^j)/O$, where the sum is over the different classes of the given feature (amino acids, secondary structure elements, and exposure class), p^i_a is the probability of observing that given feature class a in protein i, and O is the odds value definition a background score for a given feature. The log-likelihood odds values, relative weights on the three parts of the alignment function as well as the two affine gap-penalty values were optimized using a set of structurally superimposable sequence pairs with low mutual sequence similarity. Relating a sequence alignment score to a likelihood of the two sequences been structurally similar is not

straightforward. The protein length and protein amino acids composition among other things determine how a protein sequence will score against other protein sequences. We design a double-sided baseline corrected scoring scheme to allow for a direct interpretation of the alignment scoring values in terms of structural similarity likelihood. Each sequence is aligned against a set of 1500 sequence representatives with internal low sequence similarity and broad structural diversity. A baseline correction for the sequence is estimated from a least square fit of the alignment scores to the logarithm of the template query sequence. Next, a mean score and standard deviation is estimated from the baseline correction score distribution after removal of outliers. The baseline fit, mean score and standard deviation values for the two sequences are next used to determine the significance of a given alignment score. This significance score is calculated as $Z = \frac{2 \cdot Z_Q \cdot Z_T}{Z_Q + Z_T}$, where Z_Q and Z_T are the baseline corrected Z-score values for the alignment score for the query (Q) and template (T) sequences, respectively. A curated version of the PDB where the SEQRES sequence was aligned to the PDB sequence with atom coordinates was used as template database. Sequence profiles were generated using PSI-Blast with default parameters for three iterations and an e-value cut-off of 0.001[8]. Large scale benchmarking and cross validation demonstrates that the use of local structure predictions to guide the pairwise sequence alignment significantly improved the alignment quality beyond that obtained using sequence profiles only. Further, the use of double-sided baseline correction improved the specificity of the method for template recognition.

1. Rost, B. and C. Sander, Improved prediction of protein secondary structure by use of sequence profiles and neural networks. Proc Natl Acad Sci U S A, 1993. 90(16): p. 7558-62.
2. Soding, J., A. Biegert, and A.N. Lupas, The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res, 2005. 33(Web Server issue): p. W244-8.
3. Bennett-Lovsey, R.M., et al., Exploring the extremes of sequence/structure space with ensemble fold recognition in the program Phyre. Proteins, 2008. 70(3): p. 611-25.
4. Jaroszewski, L., L. Rychlewski, and A. Godzik, Improving the quality of twilight-zone alignments. Protein Sci, 2000. 9(8): p. 1487-96.
5. Petersen, T.N., et al., Prediction of protein secondary structure at 80% accuracy. Proteins., 2000. 41: p. 17--20.
6. Dor, O. and Y. Zhou, Real-SPINE: an integrated system of neural networks for real-value prediction of protein structural properties. Proteins, 2007. 68(1): p. 76-81.
7. Petersen, B., et al., NetSurfP - Predicting real value Relative Solvent Accessibility with a Pearson Correlation Coefficient of 0.70, and direct reliability predictions. In preparation, 2008.
8. Altschul, S.F., et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res., 1997. 25: p. 3389--3402.

DBAKER

Comparative Modeling of Protein Structures in CASP8 Using Full-Atom Rosetta Refinement and Manual Alignment Selection

J. Thompson[1], M.D, Tyka[1], R. Sadreyev[2], S. Raman[1], E. Kellogg[1], J. Pei[2], O. Lange[1], L. Kinch[2], B. Kim[2], F. DiMaio[1], R. Vernon[1], W. Sheffler[1], P. Barth[1], I. Davis[1],R. Das, N. Grishin[2] and D. Baker[1]

[1] – *University of Washington,*

[2] – *University of Texas Southwestern Medical Center*

dabaker@u.washington.edu

New features in our comparative modeling methodology in CASP8 include improvements to our full-atom energy function, expert-based alignment selection, and increased computational sampling. We also experimented with a variety of new protocols on subsets of targets, including the use of game player based optimization with FoldIt (http://www.fold.it/).

Alignment Methodology: We used a variety of methods to generate alignments of the query sequence to similar template structures[3,4,5]. Human experts curated the output of these programs to find high-quality alignments to the query sequence, and annotated regions of the template structure that were unlikely to be conserved in the query sequence. Using these alignments, we created starting models by copying coordinates of the aligned regions of the template structure into the query sequence.

Loop Relax: We stochastically rebuilt sections of models that were either unaligned to the template sequence, aligned with low confidence, or structurally variable within the clustered population of starting models using fragment-based modeling as in CASP7[1,2], and the resulting models were subjected to the Rosetta full atom refinement protocol. During this step all atoms are represented explicitly, and the

backbone and sidechain torsion angles of all residues are optimized using the Rosetta Monte Carlo plus minimization method described in ref 1.

Evolutionary Optimization: The process of clustering models, building loops, and minimizing the Rosetta full-atom energy was iterated several times to produce a population of models with very low Rosetta energies[1]. Submitted models were selected based on energy, visual inspection, and similarity to template structures over portions of the alignment likely to be conserved.

Results: In a number of cases, the full atom refinement led to models improved over the starting templates, including targets T0492 (Figure 1). Failures resulted from selection of incorrect alignments and overly aggressive full atom refinement, particularly in the close homology regime. As in CASP7, the accuracy of the models relative to the automated servers (which rely primarily on evolutionary information) increased with increasing target difficulty; this is likely because accurate modeling of the physical chemistry becomes more important as evolutionary information becomes weaker. We are currently working to unify, rigorously benchmark and completely automate the somewhat chaotic collection of methods we used in CASP8.

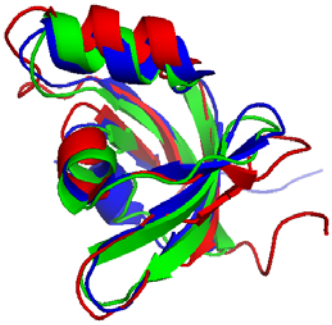


Figure 1: Superposition of native structure (blue), the best available template (red), and our best submitted model (green) for CASP8 target T0492.

1. Qian B., Raman S., Das R., Bradley P., McCoy A.J., Read R.J., Baker D. (2007) High-resolution structure prediction and the crystallographic phase problem.
2. Simons K.T., Kooperberg C., Huang E., Baker D. (1997) Assemble of Protein Tertiary Structures From Fragments with Similar Local Sequences using Simulated Annealing and Bayesian Scoring Functions. J. Mol. Biol. 268, 209-225.
3. 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.
4. Pei J., Grishin N. (2007) PROMALS: towards accurate multiple alignment of distantly related proteins. Bioinformatics 23, 802-808.
5. Söding J. (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics 21, 951-960.

DBAKER

Free Modeling of Protein Structures in CASP8 using Rosetta

R. Vernon[1], J. Thompson[1], E. Kellogg[1], W. Sheffler[1], S. Raman[1], M.D, Tyka[1], O. Lange[1], R. Sadreyev[2], J. Pei[2], L. Kinch[2], B. Kim[2], R. Das[1], N. Grishin[2] and D. Baker[1]

[1] – *University of Washington,*

[2] – *University of Texas Southwestern Medical Center*

dabaker@u.washington.edu

During CASP8 we applied the Rosetta free modeling protocol to targets for which there was no close homologue of known structure. The protocol consists of a coarse grained fragment based search of conformational space followed by physically realistic full atom refinement. Tens of thousands of independent trajectories were carried out using Rosetta@home.

Methods: The initial fragment based structure assembly step generates a diverse pool of 105-106 decoys that have buried hydrophobic cores and other protein-like features. This initial stage of modeling is carried out using a centroid level representation of the protein backbone and a low-resolution energy function1. This is followed by full atom refinement using Monte Carlo plus minimization with the Rosetta full-atom

energy function to access nearby free energy minima and make possible the recognition of the most accurate models based on their energies. The final submissions were selected by clustering the lowest energy structures, occasionally supplemented by visual inspection. In accordance with our protocols during CASP7, we increased the diversity of our models by folding multiple sequence homologs for each target3, by stochastically disallowing beta hairpins, and by resampling long-range beta sheet pairings4.

Improvements: Several improvements to our protocols were tested in CASP8, including improvements both to our energy function and our sampling strategy. Energy function improvements include an updated full atom energy function with a differentiable environment term1, and a more effective weighting of the different energy function parameters. We modified our sampling strategy in three ways. First, we generated more diverse sets at the low resolution stage by using variable fragment lengths to initially assemble structures. Second, the full folding protocol was carried out on alternative domain parses, and the results used to select the most likely parses. Third, a subset of the targets were further refined using the iterative evolutionary protocol we have used for template-based modeling4. For targets with very remote homologues of known structure, both the free modeling and template-based protocols were used and submissions were chosen from the lowest energy models overall.

Results: High-resolution predictions were made for multiple single domain targets, including all-beta (T0467; 2.5 Å over 73 residues) and alpha-beta (T0482; 2.4 Å over 82 residues) proteins. The protocol was less successful on large targets, multi-domain targets, targets with uncertain secondary structure predictions, and targets with extensive disordered regions.

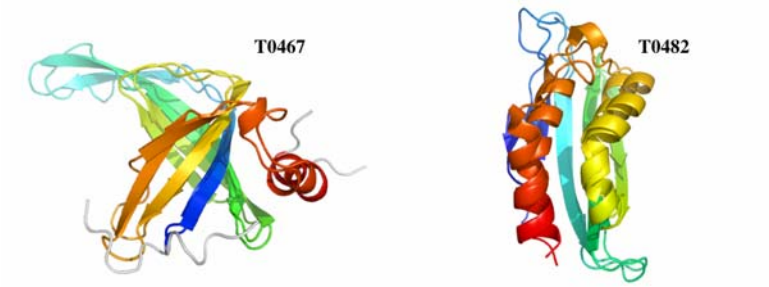


Figure 1: Predictions from the Rosetta free modeling method superimposed with NMR model 1.

1. Simons K.T., Kooperberg C., Huang E., Baker D. (1997) Assemble of Protein Tertiary Structures From Fragments with Similar Local Sequences using Simulated Annealing and Bayesian Scoring Functions. J. Mol. Biol. 268, 209-225.
2. Qian B., Raman S., Das R., Bradley P., McCoy A.J., Read R.J., Baker D. (2007) High-resolution structure prediction and the crystallographic phase problem.
3. Bradley P., Baker D. (2006) Improved beta-protein structure prediction by multilevel optimization of nonlocal strand pairings and local backbone conformation. Proteins 65, 922-9.
4. Das R, Qian B, Raman S, Vernon R, Thompson J, Bradley P, Khare S, Tyka MD, Bhat D, Chivian D, Kim DE, Sheffler WH, Malmström L, Wollacott AM, Wang C, Andre I, Baker D. (2007) Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home, Proteins 69 Suppl 8, 118-28.

Dill-ZAM

Physics-Based Protein Folding: Exploiting Locality in Folding

J.L. MacCallum[1], V.A. Voelz[2], T. Glembo[3], V.S. Pande[2], S.B. Ozkan[3], and K.A. Dill[1]
[1] – Department of Pharmaceutical Chemistry, University of California San Francisco, [2] – Department of Chemistry, Stanford University,
[3] – Department of Physics, Arizona State University
jlmaccal@maxwell.ucsf.edu

Our approach to structure prediction uses physics-based modeling. Rather than using database information – such as secondary structure predictions, template models, or PDB-derived potentials – we use a standard force field potential (Amber 96 + GBSA) in combination with molecular dynamics (MD) simulations and 'mechanism-based' sampling.

Our search strategy attempts to exploit the hierarchical nature of protein folding landscapes. Using extensive MD simulations, we search for small (<16 residue) fragments of the structure that have transiently stable secondary structures. Once identified, these fragments are assembled into full-length structures using two different methods (described below). Finally, the resulting structures are used to seed a reservoir replica exchange molecular dynamics simulation (REMD), which sorts the structures according to their free energy.

Our ultimate aims are to: (1) explore the shape and features of the protein folding landscape, (2) gain an understanding of the local nature of protein folding, (3) gauge the utility of a variety of cheap metrics for structure prediction, and (4) assess the current state of physics based force fields for protein structure prediction.

Due to the computational expense of physics based simulations, we faced severe time and resource constraints that ultimately prevented us from applying our full procedure to any target. However, our full procedure is detailed below with comments indicating steps where sacrifices were made in order to meet CASP deadlines.

Our approach works as follows:

1. We conducted extensive simulations of all 8-, 12-, and 16-residue fragments of the target sequence using the Amber 96[1] force field and an implicit solvent model[2]. These simulations were performed on the Folding@Home distributed computing network[3].
2. We identified fragments that show a propensity to form transiently stable secondary structures using residue-residue contact maps derived from the simulations.
3. Different combinations of stable secondary structures were then passed to the assembly tools described below. Due to time and resource constraints, we were typically only able to evaluate three or four of the many possible combinations of fragments.
4. The fragments were assembled using one of the following two protocols:
    a. FRODA[4] is a search algorithm based on geometric constraints. The protein is parsed into rigid units according to the bonds, hydrogen bonds, and hydrophobic contacts. These rigid units are randomly perturbed and the structure is energy minimized to satisfy the constraints. We add a metropolis routine that favors structures with a low hydrophobic radius of gyration in order to eliminate non-compact structures. We also identify pairs of hydrophobic residues and add a perturbation to push those pairs together. If we perform multiple runs and choose different pairs for each run, we then ensure unique conformations for each run and greater sampling. The resulting structures are clustered together based upon Cα RMSD. These conformations are then passed on the REMD simulations in step 5.
    b. Alternatively, for some targets we employed a rigid-body Monte Carlo sampling procedure utilizing fast loop closure routines[5]. Here, the fragments were held internally rigid and only their relative positions and orientations, and the intervening loop configurations were allowed to vary. The ensemble of structures was then filtered by radius of gyration to eliminate the non-compact structures. The remaining structures were energy minimized. Any structures with high energy were eliminated, as these typically had steric clashes or other defects. The resulting structures were then subjected to 5 ps of molecular dynamics simulations and then sorted according to average potential energy. We were typically able to generate, filter, and score 250,000 candidate structures within the timframe of CASP. Further analysis has shown that for a 100-residue protein, we would need to evaluate approximately 10^6-10^7 structures in order to identify a structure < 5Å RMSD from native.
5. The final step is to seed an REMD simulation[6] with the structures generated by step 4. REMD employs a set of parallel simulations across a range of temperatures with periodic Monte Carlo exchange of structures between adjacent temperatures. The structures with the lowest apparent free energy are identified by clustering the trajectory at the lowest temperature. Due to time constraints, this step was omitted for most targets. Instead, we typically relied on the ranking provided by the cheap metrics and sometimes visual inspection of the structures.

Our long-term goal is to extend this method to work in situations where current state of the art methods struggle due to the lack of available templates, such as for membrane proteins or peptoids, non-biological peptide mimics for which there is no database information.

1. Cornell,W.D., et al., (1995) A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **117**, 5179-5197.
2. Onufriev,A., Bashford,D. & Case,D.A. (2004) Exploring Protein Native States and Large-Scale Conformational Changes with a Modified Generalized Born Model. *Proteins: Structure, Function, and Bioinformatics*. **55**, 383-394.
3. Shirts,M.R. & Pande,V.S. (2000) Screen savers of the world, Unite! *Science* **290**, 1903-1904.

4. Wells,S., Menor,S., Hesphenheide,B. & Thorpe,M.F. (2005) Constrained geometric simulation of diffusive motion in proteins. *Physical Biol.* **2**, S127-136.
5. Coutsias,E.A., Seok,C., Jacobson,M.P. & Dill,K.A. (2004) A kinematic view of loop closure. *J. Comp. Chem.* **25**, 510-528.
6. Roitberg,A.E., Okur,A. & Simmerling,C., (2007). Coupling of Replica Exchange Simulations to a Non-Boltzmann Structure Reservoir. *Phys. Chem. Letters B* **111**, 2415-2418.

DISOclust

Intrinsic disorder prediction using the DISOclust server

L.J. McGuffin
School of Biological Sciences, University of Reading, Whiteknights, Reading RG6 6AS, UK
l.j.mcguffin@reading.ac.uk

DISOclust[1] is an unsupervised method composed of two steps; the prediction of the per-residue error in multiple fold recognition models, using ModFOLDclust[2], followed by a simple analysis of the conservation of the per-residue error across all models. The premise of the method is that residues that are highly variable in 3D space from one model to the next may coincide with regions of disorder. The DISOclust web server initially obtains multiple 3D models from the nFOLD3 server and then combines the results obtained from running the DISOclust method with those from the DISOPRED[3] method, in order to form disorder predictions for each target.

The per-residue error in each model was calculated using a score based on the average *S*-score[4]. Pairwise superpositions of the nFOLD3 server models were carried out in order to evaluate the local structural conservation of each residue in each model. For each CASP8 target, a number of nFOLD3 server models were available (N). The per-residue quality of each model was calculated by carrying out structural alignments with every other model using the TM-score program[5], with the "-d" option set to 3.9Å. In a pairwise structural alignment, if the overall TM-score was found to be >0.2, then *S*-scores were calculated for each residue (*i*) in the model. If a pair of residues were structurally aligned within the TMscore distance cut-off (as indicated by a ":" in the TM-score alignment output), then the S-score was calculated as below:

$$S_i = \frac{1}{1+\left(\frac{d_i}{d_0}\right)^2}$$

Where d_i was the distance between aligned residues and d_0 was the distance threshold (3.9). Unaligned residues in the model were given an S_i score of 0.

$$S_r = \frac{1}{N-1}\sum_{a \in A} S_{ia}$$

The mean *S*-score (S_r) was then calculated for each residue (r) in the target sequence:
Where S_r was the predicted residue accuracy for the model[5], N-1 was the number of pairwise structural alignments carried out for that model, A was the set of alignments and S_{ia} was the S_i score for a residue in a structural alignment (a). An S_r score of 0 was given to any residues that were missing from the model, so that all residues in the target sequence were scored. Finally, the DISOclust score for each residue in a target sequence was calculated as 1 minus the per-residue accuracy across all models:

$$P_d = 1 - \left(\frac{1}{N}\sum_{m \in M} S_{rm}\right)$$

Where P_d was the approximate posterior probability of a residue being in a disordered state, N was the number of models, M was the set of models and S_{rm} was the S_r score for a model (m). The P_d score for each residue was combined with the score obtained from the second column of the DISOPRED output file. Each DISOPRED score was re-scaled using $D_s=D/0.052*0.5$, for scores $\leq$0.052, and using $D_s=(D-0.052)/0.948*0.5+0.5$, for scores >0.052, where D was the original DISOPRED score and D_s was the rescaled value. The mean of the P_d and D_s was taken as the probability of disorder for each residue.

The DISOclust web server is available at the following URL:
http://www.reading.ac.uk/bioinf/DISOclust/

1. McGuffin,L.J. (2008) Intrinsic disorder prediction from the analysis of multiple protein fold recognition models. *Bioinformatics*. **24**, 1798-1804.
2. McGuffin,L.J. (2008) The ModFOLD Server for the Quality Assessment of Protein Structural Models. *Bioinformatics*. **24**, 586-587.
3. 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.
4. Levitt,M. & Gerstein,M. (1998) A unified statistical framework for sequence comparison and structure comparison. *Proc Natl Acad Sci U S A*. **95**, 5913-5920.
5. Zhang,Y. & Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins*. **57**, 702-710.

Distill

Distill, Shandy, Punch: draft protein structures by machine learning

D. Baù[1], A. J. Martin[1], C. Mooney[1], A. Vullo[1], I. Walsh[1] and G. Pollastri[1]*
[1] - *School of Computer Science and Informatics,*
University College Dublin, Ireland
gianluca.pollastri@ucd.ie

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

Secondary structure is predicted by Porter[1], relative solvent accessibility by PaleAle[2], contact density by BrownAle[3], residue contact and distance maps by XXStout[3]. Residue contact maps submitted to CASP (8Å) are obtained by XXStout, and are not directly used to predict 3D coordinates. 4-class distance map predictions by an architecture identical to XXStout's are adopted instead. All structural feature predictors are based on single- or dual-layer Recursive Neural Network architectures for Directed Acyclic Graphs (DAG RNNs)[4]. One-dimensional feature predictors (i.e. those mapping the primary sequence into a sequence of the same length) are based on 1D DAG RNNs, while contact and distance map predictors are based on 2D DAG RNNs. All predictors are provided structural information about PDB templates as a further input, when templates are available. Templates are identified as follows: 2 rounds of PSI-BLAST are run against UniProt; the resulting PSSM, plus predictions of structural motifs by Porter+[5], are aligned locally against all the sequences and corresponding structural motifs in the PDB. If no suitable template is found this way (e-value<1e-3), we predict all 1-dimensional properties ab initio, and search for templates by alignigning the complex of these properties plus the sequence against the equivalent representation of all PDB proteins, by dynamic programming.

In the next stage, we reconstruct sets of Cα coordinates. The reconstruction is carried out by minimising a potential function containing terms that penalise the violation of predicted distances between residues, and enforce predicted strand locations, hard-core repulsion between amino acids, and virtual Cα-Cα bond lengths. The actual search is performed in 3 stages:

- Initial structures are generated, in which helices predicted by Porter are modelled, consecutive Cα atoms are set at a realistic distance (~3.8Å), and virtual Cα angles are restricted to the 90º-180º interval.
- A search from these initial structures is performed by introducing perturbations in them. 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. 5,000 moves of every non-helical Cα and helical termini are attempted for each search. 50 searches are run for each protein structure.
- Finally, the structures obtained are ranked. In the *ab initio* case we rank the structures by a neural network trained to map a number of characteristics (enforcement of of predicted constraints, secondary structure composition, compaction, etc.) of each structure into its quality, measured as its TM score against the correct structure. In the case templates from the PDB are available, similarity to the templates is used as further information for ranking.
- Finally, all atom models are obtained by reconstructing the backbone via maxsprout[7] and the side chains by SCWRL[8].

We also submitted predictions of protein domains and protein disorder by predictors that are not integrated in Distill's pipeline. The predictor or protein domains (Shandy) has three stages: one in which SCOP and PDB templates are found; a second stage (a 1D DAG Recurrent Neural Network) in which residues are marked as domain boundary vs. intra-domain using primary and template information; a third stage in

which the previous predictions are smoothed and the location of domain boundaries is decided. Disorder is predicted by Punch, an evolution of Spritz[6], a combination of experts implemented by kernel machines, which also uses template information.

1. Pollastri,G. & McLysaght,A. (2005) Porter, A new, accurate server for protein secondary structure prediction, Bioinformatics, 21(8), 1719–1720.
2. Baù,D., Martin,A.J.M., Mooney,C., Vullo,A., Walsh,I. & Pollastri,G. (2006) Distill: A suite of web servers for the prediction of one-, two- and three-dimensional structural features of proteins, BMC Bioinformatics, 7, 402.
3. Vullo,A., Walsh,I. & Pollastri,G. (2006) A two-stage approach for improved prediction of residue contact maps, BMC Bioinformatics, 7, 180.
4. Baldi,P. & Pollastri,G. (2003) The Principled Design of Large-Scale Recursive Neural Network Architectures – DAG-RNNs and the Protein Structure Prediction Problem.
5. Journal of Machine Learning Research, 4, 575-602.
6. Mooney,C., Vullo, A. & Pollastri, G.. (2006) Protein Structural Motif Prediction in Multidimensional φ-ψ Space leads to improved Secondary Structure Prediction, Journal of Computational Biology, 13(8), 1489-1502.
7. Vullo,A., Bortolami,O., Pollastri,G. & Tosatto,S. (2006) Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines, Nucleic Acids Research, 34, W164-W168.
8. 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(1):183-94.
9. Canutescu AA, Shelenkov AA, Dunbrack RL Jr.(2003) A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci., 12(9):2001-14.

DISTILLF

A.J. Martin and G. Pollastri
University College Dublin
gianluca.pollastri@ucd.ie

DISTILLF is a new knowledge-based Model Quality Assessment Program (MQAP) at the residue level which evaluates single protein structure models. DISTILLF also predicts local quality, but it is derived from global quality. In each structure model each Amino Acid (AA) is represented by its C-alpha. We consider two AAs as interacting if their C-alphas are at up to a distance of 20┤ A. Each interaction between AAs is evaluated individually by a Neural Network (NN), which produces a vector of hidden features as output. The features from all interacting pairs are obtained from as many copies of the NN as there are interactions, then added up and presented to a further NN which maps the resulting vector into a measure of the global goodness of the structure/decoy. The whole, compound network (all the interaction network copies plus the output network) is trained by backpropagating the difference between global goodness and actual model quality. As target function we use TMScore as it is a model quality measurement independent of the model length and more sensitive to details than GDT TS or RMSD[1]. To train the NN we used models submitted to CASP editions 5, 6 and 7 in 5 fold cross-validation. Values stored in the hidden states after representing each AA correlate with the scaled distance used in the TMScore calculation (local quality measurement). As inputs for the NN we use a vector of numbers that describes each pair of AAs and their interaction. This input vector contains several structure descriptors computed solely from the C-alpha trace. These structure descriptors encode each AA's environment, the interaction between two AAs in contact and their identities. AAs environment is described by distances with sequence neighbours, several angles formed between the AA's C-alpha and C-alphas of its sequence neighbours, pseudo solvent accessibility as HSE measure[2], pseudo packing quality, angles of HSE's pseudo C-beta vectors with sequence neighbours' pseudo C-betas. The interaction between two AAs in contact is described by the distance of each AA in the pair and its sequence neighbours to the other AA of the pair and its sequence neighbours, and the angles between their respective pseudo C-beta vectors. The AAs identities are also provided to the network.

1. Zhang, Y. & Skolnick, J. (2004) Scoring Function for Automated Assessment of Protein Structure Template Quality. *PROTEINS: Structure, Function, and Bioinformatics* **57**, 702-710.
2. Hamelryck, T. (2005) An Amino Acid Has Two Sides: A New 2D Measure Provides a Different View of Solvent Exposure, PROTEINS: Structure, Function, and Bioinformatics, 59, 38-48.

Structure evaluation program using the local consensus-based similarity and circle quality assessment method

G. Terashi[1], H. Sakai[1], K. Kanou[1], T. Hirata[1], M.Takeda-Shitaka[1], and H. Umeyama[1]
[1] - *School of Pharmacy, Kitasato University*
terashig@pharm.kitasato-u.ac.jp

In the CASP8, our fams-ace2 server participated in the 3D coordinate prediction category as a human expert group. We applied two different scoring functions for the fully automated model prediction server, fams-ace2: (1) the local consensus score; and (2) the model quality score based on classification of the side-chain environment for each residue. The local consensus score was used as a filter to select the models which have locally similar structures comparing with the set of models. The model quality score was then used for the final selection of the best model. This model quality score was calculated by our model quality assessment program CIRCLE[1].

The procedure of fams-ace2 can be summarized as the following 4 steps:

$$LocatCons_m = \frac{\sum_{n}^{N}\sum_{i}^{R_m} MAXSUB(LOC_{m,i}, LOC_{n,i})}{N} \quad (1)$$

(1) Obviously incorrect models which have serious physical clashes or broken main-chain structures were removed. (2) The top 10% (an optimized parameter of fams-ace2) of server models were selected in the order of the local consensus score; the local consensus score is calculated as the equation (1). *N* is the number of server models. *LOCm,i* is a set of C-alpha coordinates which exist within 10Å from the i th residue of model *m*. *MAXSUB(a,b)* is a maximum number of C-alpha coordinates (subset *a*) which superimpose well (within 3Å) upon their corresponding C-alpha coordinates in subset b. The values of 10 and 3 Å are optimized parameters of fams-ace2. (3) All of the server models, selected in step (2), were refined and rebuilt utilizing our homology modeling program FAMS[2]. (4) The top 5 structures were selected, according to a model quality evaluation based on their CIRCLE score. The coefficients of *SSscore* in the circle which do not use the consensus method were changed in the fams-ace2 from 0.35 and 0.75 to 0.30 and 0.30, respectively. The fams-ace2 is a fully automated server and does not require human intervention.

The parameters of fams-ace2 were optimized by the data set of previous CASP7. We used the GDT_TS as the quality of model compared to native. When we applied optimized fams-ace2 to CASP7 targets, fams-ace2 obtained the best results over all server groups (Fig.1). Moreover, in Template Based Modeling Targets, fams-ace2 also achieved best results over all groups including human groups.

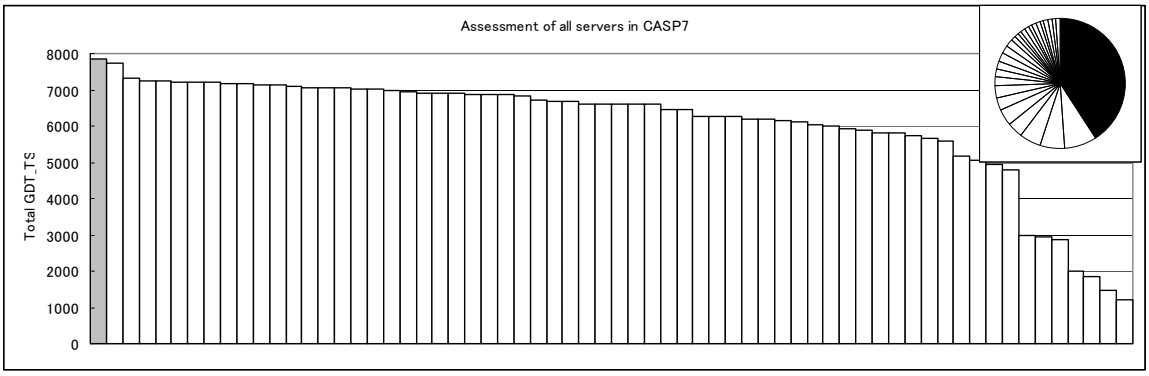


Fig. 1 Results of fams-ace2 (gray bar) and distribution of selected servers (pie graph) in CASP7

The 103 native protein structures of CASP8 128 targets were published in CASP8 web site (Sep 03 2008). We calculated GDT_TS of all models submitted by servers and fams-ace2 (Fig. 2). The total GDT_TS of fams-ace2 (gray in Fig.2) were obviously better than almost all of the other servers. The fams-ace2 selected models of the best server (Zhang-server) among 40% and 34% of targets in CASP7 and CASP8, respectively (The black area in pie graph of Fig.1 and Fig.2).

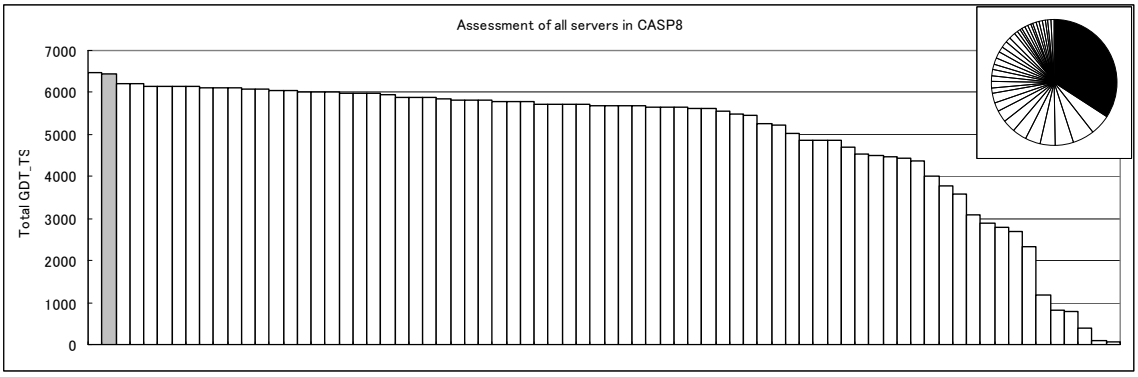


Fig. 2 Results of fams-ace2 (gray bar) and distribution of selected servers (pie graph) in CASP8

Although the advantage of fams-ace2 over other servers is slightly smaller than the results applying for CASP7 (123 domains, Fig.1), the intended results are accomplished. This small difference between CASP7 and CASP8 might be caused by the change of the distribution of target difficulty and performance of servers. When we calculate GDT_TS of CASP8 models, we did not consider the domain regions. Therefore the results of some targets will be changed. The advantages of fams-ace2 are the fully automated process, the lower calculation costs due to the decrease of the modeled number in comparison with Fams-ace[1], and a high accuracy similar to the top of human groups. We are planning to optimize fams-ace2 according to the target difficulty and performance of each server by using much huger data set.

1. Terashi G, Takeda-Shitaka M, Kanou K, Iwadate M, Takaya D, Hosoi A, Ohta K, and Umeyama H (2007). Fams-ace: a combined method to select the best model after remodeling all server models. *Proteins*.69 Suppl 8:98-107.
2. Ogata, K. and Umeyama, H. (2000) An automatic homology modeling method consisting of database searches and simulated annealing. *J. Mol. Graphics Mod*. 18(3):258-72, 305-6.

DomFOLD

Automated protein domain prediction using the DomFOLD server

L.J. McGuffin[1]
[1] – School of Biological Sciences, University of Reading, Whiteknights, Reading RG6 6AS, UK
l.j.mcguffin@reading.ac.uk

The DomFOLD server uses a consensus of three different methods for domain prediction. The output from DomSSEA[1], HHsearch[2], and DISOPRED[3] is parsed to form a domain prediction for each method. The final prediction is then a simple majority vote taken on the domain assignment of each residue. Where the vote is evenly split, the lowest domain number is taken.

The first method used for domain prediction is DomSSEA, which has been described previously[1]. DomSSEA is based on the alignment of the PSIPRED[4] predicted secondary structure of the target against a fold library of known secondary structures, determined using DSSP[5]. The domain boundaries of templates within the fold library are assigned using SCOP[6], which are then mapped onto the target structure.

The second method parses the top alignments from HHsearch. Domain boundaries are assigned by the location of each template aligned to the target sequence. Where possible, the boundaries of aligned templates with multiple domains are appropriately subdivided using the SCOP domain assignment. The consensus domain assignment is then used to determine the overall domain boundaries for this method.

The third method is based on disordered regions predicted using the DISOPRED method. The premise of this method is that regions of the target protein that are predicted to be disordered may indicate flexible domain linkers. Domain boundaries are predicted in stretches of disorder which are more than twenty residues from the N- and C-termini.

The DomFOLD web server is available at the following URL:
http://www.reading.ac.uk/bioinf/DomFOLD/

1. Marsden,R., McGuffin,L.J. & Jones,D.T. (2002) Rapid protein domain assignment from amino acid sequence using predicted secondary structure. *Protein Sci.* **11**, 2814-2824.

2. Söding J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*. **21**, 951-996.
3. 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.
4. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol.* **292**, 195-202.
5. Kabsch,W. & Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers.* **22**, 2577-2637.
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.

DOMSERV_H&E

Domain prediction using protein structure prediction and improved DLP-SVM

T. Ebina[1] and S. Hirose[1,2]

[1] – *Department of Biotechnology and Life Science, Tokyo University of Agriculture and Technology (TUAT), Japan*

[2] – *Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology, Japan*

teppei-ebina@nifty.com

We have developed a domain prediction method based on the results of inter-domain loop (domain linker[1]) prediction, in combination with those of various protein structure prediction tools including fold recognition. Our method consists of two alternative prediction schemes, one of which is selected, depending on whether structures fit to a target sequence can be detected.

First, we executed well known fold recognition methods, HHsearch[2] and GenThreader[3], to detect similar structure for target sequence. From the results of fold recognition, proteins with scores higher than 90 (HHsearch) and/or 4.0 (GenThreader) were selected as template structures of target sequence. If template structures were detected, the domain regions of target sequences were manually determined based on the template structures. Even when only a part of protein structures were selected as the template, we carefully assigned domain regions to the target, referring the template structure.

When no template structure was detected, we then determined domain regions using the prediction results of domain linker candidates, in combination with those of secondary structure and disordered regions. In this scheme, domain linker candidates were predicted by an improved version of DLP-SVM[1] that uses position specific scoring matrix (PSSM) for its training and prediction. The output values of DLP-SVM represent the domain linker propensity of each residue. Thus we predicted domain linker candidates based on the raw output values. To reduce the false positives, we refine the domain linker candidates using the prediction results of disordered regions and those of secondary structure by POODLE[4-6] series and PSIPRED[7], respectively. Since POODLE series individually predicted the different types of disordered regions, we totally assessed the prediction results and determined disordered regions within the target sequences. In the next step of this scheme, we selected domain linker candidates with significantly higher DLP-SVM output values than those of other regions or with strongly predicted as disordered region and/or coil regions, or with both. According to selected domain linkers, we determined the domain regions, where existing the both side of the selected region. The results of these protein structure prediction and domain assignments were carefully analyzed and final domain regions were determined.

1. Ebina T.,Toh H. & Kuroda Y. (2008) Loop-length dependent domain linker prediction for high-throughput structural proteomics, *Peptide Science*, in press.
2. Söding J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21(7)**, 951-960.
3. Jones DT. (1999). GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol*., **287(4)**, 797-815.
4. Shimizu K., Hirose S. & Noguchi T. (2007). POODLE-S: web application for predicting protein disorder by using physicochemical features and reduced amino acid set of a protein-specific scoring matrix. *Bioinformatics*, **23(17)**, 2337-2338.
5. Hirose S., Shimizu K., Kanai S., Kuroda Y. & Noguchi T. (2007). POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions. *Bioinformatics*, **23(17)**, 2046-2053.
6. Shimizu K., Muraoka Y., Hirose S., Tomii K. & Noguchi T. (2007). Predicting mostly disordered proteins by using structure-unknown protein data. *BMC Bioinformatics*, **8**, 78.
7. McGuffin J., Bryson K. & Jones DT. (2000). The PSIPERD protein structure prediction server. *Bioinformatics*, **16(4)**, 404-405.

Elofsson

Prediction of A2a receptor as step forward automatized pipeline for GPCR-ligand complex prediction.

W. Jurkowski[1] and A.Elofsson[1]
[1] - *Center of Biomembrane Research, Department of Biochemistry & Biophysics, Stockholm University*
jurkow@sbc.su.se

Procedure to automatize the GPCR structure prediction and ligand docking is presented. The aim of method under development is to use common tools (e.g. Modeller[1], Autodock[2]) and to limit human intervention on each of modeling phase: homology modeling of receptor, ligand preparation, binding site determination, docking and scoring. Recently published structure of Adenosine Receptor A2a complexed with antagonist ZM241385[3], which could be blind predicted thanks to Critical Assessment of GPCR Modeling and Docking 2008 project[4] served as one of the training targets. Results of this experiment are with good agreement with the crystal structure with less than 1.5 Å RMSD for TM and external loops and correct pose of the ligand. Even tough, the prediction is not faultless with ligand to much buried inside the TM part of protein it shows positive perspective for this prediction routine.

1. N. Eswar, M. A. Marti-Renom, B. Webb, M. S. Madhusudhan, D. Eramian, M. Shen, U. Pieper, A. Sali. Comparative Protein Structure Modeling With MODELLER. Current Protocols in Bioinformatics, John Wiley & Sons, Inc., Supplement 15, 5.6.1-5.6.30, 200.
2. Morris, G. M., Goodsell, D. S., Halliday, R.S., Huey, R., Hart, W. E., Belew, R. K. and Olson, A. J. (1998), Automated Docking Using a Lamarckian Genetic Algorithm and and Empirical Binding Free Energy Function J. Computational Chemistry, 19: 1639-1662.
3. Jaakola VP, Griffith MT, Hanson MA, Cherezov V, Chien EY, Lane JR, Ijzerman AP, Stevens RC. (2008), The 2.6 Angstrom Crystal Structure of a Human A2A Adenosine Receptor Bound to an Antagonist. Science (ahead of print)
4. Brooks C., Dixon S. and Moult J. Critical Assessment of GPCR Modeling and Docking 2008,www: http://jcimpt.scripps.edu/GPCRDock/default.aspx

fais-server

Homology modeling and *de novo* structure prediction based on contact number prediction.

M. Shirota[1], T. Ishida[1], K. Kinoshita[1, 2]
[1]-*Human Genome Center, Institute of Medical Science, the University of Tokyo*
[2]-*Institute for Bioinformatics Research and Development, JST*
mshirota@hgc.jp

Fais-server is an automated protein tertiary structure and disorder prediction server.

For tertiary structure prediction, the server tried to identify the structural templates of a target sequence for the PDB library by fold recognition technique. Fold recognitions were done by HMM-HMM comparison using HHsearch program[1] and by our profile-profile alignment program, which searches for templates with the similar position specific scoring matrices (PSSM) by PSI-BLAST along with the matches of secondary structures of the template and predicted ones for the target by Psipred. If good templates with statistical significance were found, tertiary models were generated with those templates by Modeller program. If there were some long unaligned regions in the alignment, those regions were modeled by our *de novo* structure prediction system described later. The models were ranked according to the statistical significance of the alignment and the best five models were submitted.

If reliable alignments could not be found for the target, the server generated tertiary structure models by *de novo* modeling system based on the fragment assembly method. Candidate fragments for each position of the target sequence were searched using the Pearson's correlation coefficient between PSSMs of query subsequences and that of the target subsequence. Using that fragment libraries, the server searched conformational spaces using a potential energy function by simulated annealing method. Our potential energy function includes terms of potential based on contact number prediction[2], atom clashes, and hydrogen bonding. About 1,000 models were produced for each target, and five prediction models were

selected by using the potential energy and structural clustering. Finally, side chain modeling was performed by using SCWRL version 3.0[3].

For the prediction of disordered protein regions, the server used our protein disorder prediction system, named PrDOS[4]. The prediction system is composed of two predictors: a predictor based on local amino acid sequence information and one based on template proteins. The first part is implemented using a support vector machine (SVM) algorithm for the PSSM of the input sequence. The second part assumes the conservation of intrinsic disorder in protein families, and is simply implemented using PSI-BLAST and our own measure of disorder. The final prediction is done as the combination of the results of the two predictors.

1. Soding J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**(7), 951-960.
2. Ishida T, Nakamura S, Shimizu K. (2006) Potential for assessing quality of protein structure based on contact number prediction. *Proteins* **64**(4): 940-947.
3. Canutescu A.A., Shelenkov A.A. & Dunbrack Jr.,R.L. (2003) A graph theory algorithm for protein side-chain prediction. *Protein Sci.* **12**, 2001-2014.
4. Ishida T and Kinoshita K. (2007) PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Res*. **35**(Web Server issue):W460-464.

fais@hgc

Model selection based on the combination of multiple energy functions and consensus of structures and function prediction based on the molecular surface of predicted structures

M. Shirota[1], T. Ishida[1], K. Kinoshita[1]

[1]-*Human Genome Center, Institute of Medical Science, the University of Tokyo*

[2]-*Institute for Bioinformatics Research and Development, JST*

mshirota@hgc.jp

Our strategy of predicting tertiary structures consists of producing models by our own *de novo* modeling methods and selecting appropriate models from those and server models. First, we classified the targets as either template based modeling (TBM) targets or free modeling (FM) targets according to their PSI-BLAST e-values. We took different approaches for each category.

For TBM targets, we did not generate tertiary structure models. We collected the server models and selected the good models according to our scoring functions. At first, we picked up the server models from CASP8 web site, in which the coordinates of all the heavy atoms were present. We evaluated these models by a scoring function that consists of multiple energy functions and the score based on the structural consensus between the models. We used three types of potential functions; potential based on the contact number prediction[1], Verify3D[2] and our statistic potentials depending on the distance between atom pairs. For the distance dependent atom pair potentials, we used four difference reference states. Two of them have the similar reference states with RAPDF[3] and DFIRE[4] and the others have hybrid reference states of the former two potentials. For each potential, the mean and standard deviation was calculated for all server models, and the potential values of each structure were transformed into Z-scores. We also used the structural consensus between target model and the other server models for scoring. For each model, we calculated the average of the GDT-TS scores from all the other server models and used it as a measure of the structural consensus. The target models were evaluated by the weighted sum of Z-scores of potentials and the structural consensus score. We selected the best five models according to the score and submitted these models after refinement by using short Monte Carlo minimization with the scoring function.

For FM targets, we generated tertiary structure models by using our *de novo* modeling system based on the fragment assembly method. Our potential energy function includes terms of potential based on contact number prediction[1], atom clashes, and hydrogen bonding. About 3,000 models were produced for each target. Sidechains were modeled by using SCWRL. Finally, we selected five structures from the server prediction models and these models as in the case of TBM targets.

For function prediction, we predicted binding sites of the target proteins based on the similarity searches of molecular surfaces against the database to representative heteroatom binding sites appearing within the Protein Data Bank (PDB) using predicted models. For searching binding sites, we used the structure of the best models selected previous process as queries and submitted them to eF-seek[5] server. Finally, we submitted the five best candidates by the server as prediction results.

1. Ishida T, Nakamura S, Shimizu K. (2006) Potential for assessing quality of protein structure based on contact number prediction. *Proteins* **64**(4), 940-947.
2. Bowie JU, Luthy R, Eisenberg D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253**, 164-170.
3. Samudrala R, Moult J. (1998) An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J Mol. Biol.* **275**(5), 895-916.
4. Zhou H, Zhou Y. (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.* **11**(11), 2714-2726.
5. Kinoshita, K., Nakamura, H. (2005) Identification of the ligand binding sites on the molecular surface of proteins. *Protein Sci.* **14**, 711-718.

FALCON

Fragment-HMM: A New Approach To Protein Structure Prediction

S. C. Li[1], D. Bu[1,2], J. Xu[3], and M. Li[1]
[1] *-David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1,* [2] *-Institute for Computing Technology, Chinese Academy of Sciences, China,* [3] *-Toyota Technological Institute at Chicago, Chicago, IL, USA, 60637*
mli@uwaterloo.ca, j3xu@tti-c.org

We wish to find a unified and the simplest model for protein structure prediction, one of the major open problems in science. We are not interested in trying PSI-BLAST for easy targets, threading by RAPTOR for harder targets, fragment assembly by ROSETTA for *ab initio* targets, or consensus for everything. We are also not interested in using different methods for different steps, such as Monte Carlo fragment assembly, clustering, selecting, refinement. Nature does not do this. It does not fit with the Occam's Razor principle. Nature prefers simplicity. We wish to find one theory, one model, as simple as possible, that goes from an input sequence to the final structure. This theory should embody homology modeling, threading, fragment assembly (all stages of it), loop modeling, refinement, side chain packing, and consensus. This theory must be simple, robust, and effective. This paper presents our initial efforts in building a theory toward this goal, and our preliminary implementation of this theory, FALCON, together with clear cut experimental results. Some ideas of our work come from three lines o research: fragment assembly, hidden Markov model sampling, and Ramachandran basins.

We propose a simple and unified paradigm for protein structure prediction. The plan is to probabilistically sample protein structure conformations compatible with local structural biases for a given protein. The architecture of the model is as below.

1. For residue i, several *Cosine* models are used to describe the local bias of its torsion angle pair (ϕ_i, ψ_i).
2. A position specific hidden Markov model (HMM) is used to capture the dependencies among local biases of adjacent residues, based on carefully selected fragments. This HMM is referred to as Fragment-HMM.
3. The Fragment-HMM is used to sample a sequence of torsion angle pairs for the given protein sequence. An energy function is used to evaluate the generated decoys, and to direct the sampling process to the better decoys.
4. The generated decoys are fed back to produce more accurate estimations of local structural biases, a more accurate Fragment-HMM and thus, better decoys. This step is executed iteratively to increase the quality of the final decoys, until convergence.

This model has advantages over existing works as follows.

─Our Fragment-HMM model combines the very successful fragment assembly method and the elegant FB5-HMM idea. Rather than using the fragments as building blocks, we use them to produce local bias information. We use the directional distribution to model local biases, and use HMM to explore the dependency among the adjacent residues. Unlike FB5-HMM, our Fragment-HMM is position specific.

─ Our Fragment-HMM naturally enables the Step 4 to re-sample decoys. Immediately, the readers would observe that this applies to obtaining fragments from a known structure. Thus this naturally enables homology modeling, threading, refinement (requiring more hidden nodes to model side chains), loop modeling, and consensus, unifying all these approaches under one roof.

─ Step 4 is similar to that of primal and dual optimization process. The primal goal is to minimize the energy which is done by discriminating decoys with an energy function; and the dual process is done via

sampling our Fragment-HMM to improve the estimation of torsion angles. Step 4 differs from the traditional fragment assembly methods that end with a population of decoys: some good and some bad. Our model does not stop here, but iterates until convergence.
–The search space is narrowed down step by step. Monte Carlo is a popular technique for fragment-assembly-based protein structure prediction. However, Monte Carlo suffers from its low efficiency since it does not explore the characteristics of the search space. In contrast, our Fragment-HMM narrows down the search space after each iteration step since the local structural biases are estimated more and more accurately.

We have implemented this theory in FALCON, Fragment-HMM approximating local bias and consensus.. We take all 6 proteins from the ROSETTA benchmark data used in (Simons *et al.*, 1997). FALCON converges 100% to within 6 Å for all six proteins after only four iterations.

FALCON was designed for short targets. During the CASP8, it was evolved to take longer target by using some of the threading results (as longer fragments) as input. However, more it is still evolving to combine the strength of threading programs.

FAMSD

Individual comparative modeling server using SP3 & SPARKS2, FAMS and CIRCLE.

K. Kanou, T. Hirata, G. Terashi, H. Sakai,
M. Takeda-Shitaka and H. Umeyama
School of Pharmacy, Kitasato University, 5-9-1 Shirokane, Minato-ku,
Tokyo 108-8641 JAPAN

Our comparative modeling method consists of following four steps: (1) making sequence alignments between target protein and template structures, (2) constructing three-dimensional structures based upon each alignment, (3) selecting the best structure model and (4) refinement of the selected model. Programs such as SP3 [1], FAMS (Full Automatic Modeling System) [2], CIRCLE [3] and Molecular dynamics were used at the each step (1) ~ (4), respectively.

(1) Making sequence alignments
8 kinds of alignment programs, BLAST, PSI-BLAST [4], PSF-BLAST, RPS-BLAST, IMPALA, Pfam-BLAST, SPARKS2 and SP3 were executed for each target protein sequence. Various alignments were generated and were filtered with its alignment score. The alignment scores for 6 kinds of methods except SPARKS2 and SP3 were calculated with following equation,

$$score = f(k_i, Hom, Len, SS) \quad (1)$$

Here *Len* is the number of residues of a predicted model. *Hom* indicates sequence identity % value, *SS* is the degree of secondary structure agreement between the secondary structures predicted one from sequence using PSI-PRED [5] and one calculated from model using STRIDE. k_i is a coefficients for each alignment method.
And as the alignment score for SPARKS2 and SP3, Z-score of their output was used. When the alignment score was more than (the maximum score of all alignments) * X, these alignments were used to construct model. A parameter X is a cut-off value which was decided using CASP7 targets as a training set depending on difficulty of each target (Table 1). The difficulty was predicted using Support Vector Machine (SVM). The alignment score and sequence identity of PSI-BLAST and these of SPARKS are used as parameters for SVM training.

(2) Constructing three-dimensional structures
We constructed three-dimensional structures using FAMS program based on each selected alignment which was mentioned in the preceding section.

(3) Selecting the best structure
All constructed models were evaluated using following scoring function,
score = *CIRCLE* + *w* * *SSscore*
Here, *Circle* represents the 3D1D score which was improved based on verify3D and *SSscore* represents the degree of secondary structure agreement. *w* is the weighting factor for *SSscore* which was optimized using CASP7 models as a training set (Table 1).

Table 1. Values of X and *w*.

PSIB	SPK2	X	*w*
CMeasy	CMeasy	**0.99**	**0.00**
CMhard	CMeasy	**0.90**	**0.30**
CMeasy	CMhard	**0.85**	**0.35**
CMhard	CMhard	**0.85**	**0.35**
FRorNF	CMhard	**0.85**	**0.35**
CMhard	FRH	**0.80**	**0.55**
CMhard	FRAorNF	**0.80**	**0.55**
FRorNF	FRH	**0.80**	**0.55**
FRorNF	FRAorNF	**0.80**	**0.55**

PSIB and SPK2 indicate the predicted difficulty using alignment score and sequence identity of PSI-BLAST and these of SPARKS2, respectively, as parameters for SVM.

Figure 1 shows the distribution of alignment method of finally ranked first models by above scoring function.

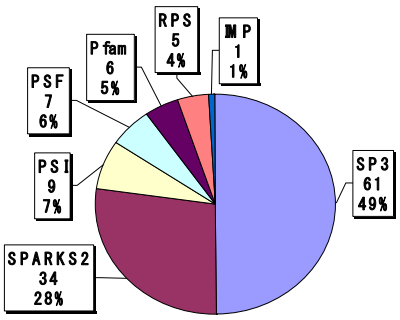


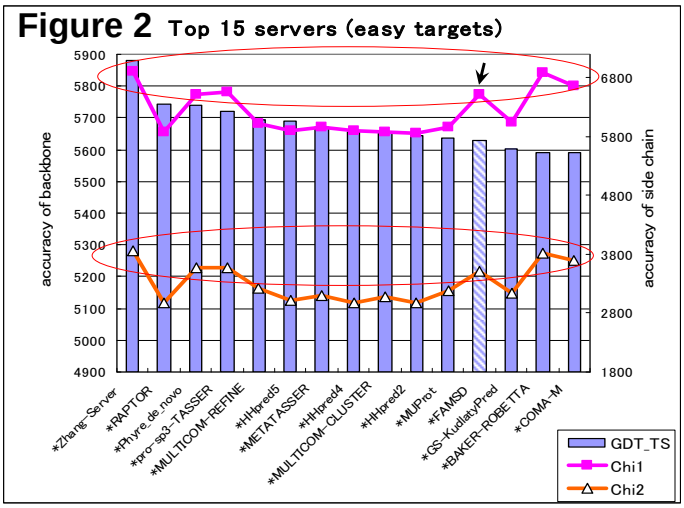
Figure 1

(4) Refinement of the selected models

Five selected models were refined using Energy minimize & Molecular dynamics. With this procedure, hydrogen bonds, main chain torsion angles and side chain torsion angles were refined slightly and collisions of hydrophobic atoms were decreased.

Results and discussion

109 experimental structures of 128 CASP8 targets became available by October 3, 2008. We evaluated the quality of all server models. FAMSD ranked at 15th with the cumulative GDT_TS score of all 109 targets. The accuracy of side chain was also assessed with the number of residues in the case that each model have a sufficiently accurate side chain, i.e., chi1 and chi2 torsion angle which is within 30 and 60 degrees, respectively, from native structure. Furthermore we calculated the cumulative score of GDT_TS, chi1 and chi2 for only 80 targets in the CM category (Figure 2). Target classification is referred to on Robetta evaluation page [6]. As the results, the rank of FAMSD with GDT_TS, chi1 and chi2 were 12th, 7th and 10th, respectively. The six servers (Zhang-Server, Phyre_de_novo, pro-sp3-TASSER, FAMSD, BAKER-ROBETTA and COMA-M) predicted high quality models in terms of not only backbone geometry but also side chain conformation.



[1] Zhou H, Zhou Y. Proteins. 2005;61 Suppl 7:152-6.
[2] Ogata, K. and Umeyama, H. J Mol Graph Model 2000; 18, 258-272.
[3] Terashi G, Takeda-Shitaka M, Kanou K, Iwadate M, Takaya D, Hosoi A, Ohta K, Umeyama H. Proteins. 2007;69 Suppl 8:98-107.
[4] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. Nucleic Acids Res. 1997; 25, 3389-3402.
[5] McGuffin LJ, Bryson K, Jones DT. Bioinformatics. 2000; Apr;16(4):404-5.
[6] http://robetta.bakerlab.org/CASP8_eval/index.html

FAMSD

FAMSD_QA: Quality assessment based on the side chain environment consensus score

K. Kanou[1], T. Hirata[1], G. Terashi[1], H. Sakai[1],
M. Takeda-Shitaka[1] and H. Umeyama[1]
School of Pharmacy, Kitasato University, 5-9-1 Shirokane, Minato-ku, Tokyo 108-8641 JAPAN
kanouk@pharm.kitasato-u.ac.jp

A consensus method like 3D-Jury [1] is one of the most powerful methods of model quality assessment. 3D-Jury score represent consensus of the backbone geometry among structure models. This method can select "good backbone" models but the quality of the side chain of selected models is not so good.
Thus we developed a new consensus method which considers side chain environment for the purpose of selecting good side chain models, and participated in Quality Assessment category using this method as a team FAMSD. We describe the algorithm of this method and our results for CASP8.

Methods
First, we calculated the side chain environment composed of 'fraction buried' and 'fraction polar' for each residue of predicted model. 'Fraction buried' is the fraction of buried area within the surrounding side chain atoms, and 'fraction polar' is the fraction of buried area within the surrounding polar atoms. These values range from 0 to 1.0 per residue. When the model A was assessed, for each residue of model A, the side chain environment was calculated and is compared with the other models. If the Euclidian distance between the side chain environment ('fraction buried' and 'fraction polar') [2] of one residue of model A and that of corresponding residue of another model was within 0.2, we considered that the two residues were in the same environment. For each model, we counted the number of residues in the same environment and the side chain environment score is the summation of those numbers. The threshold of 0.2 was determined using CASP7 models as a training set.

In CASP8, we participated in QA category as a team 'FAMSD_QA'. We had refined all predicted models by FAMS [3] and had assessed quality of these models using following combined score.

$$score = env\_con + w * SSscore$$

Here, *env_con* represents the side chain environment consensus score and *SSscore* represents the degree of match between the secondary structure of a predicted model and the secondary structure predicted from the given sequence with PSIPRED [4]. *w* is the weighting factor for *SSscore* and ranges from 0 to 1 depending on the predicted difficulty using SVM. In the case of difficult targets, more weight is given to *SSscore* than easy targets. This value was optimized using CASP7 models.

PSIB	SPK2	w
CMeasy	CMeasy	**0.3**
CMhard	CMeasy	**0.3**
CMeasy	CMhard	**0.5**
CMhard	CMhard	**0.5**
CMhard	FRH	**0.5**
FRorNF	CMhard	**0.5**
FRorNF	FRH	**1.0**
CMhard	FRAorNF	**1.0**
FRorNF	FRAorNF	**1.0**

PSIB and SPK2 indicate the predicted difficulty using alignment score and sequence identity of PSI-BLAST and these of SPARKS2, respectively, as parameters for SVM training.

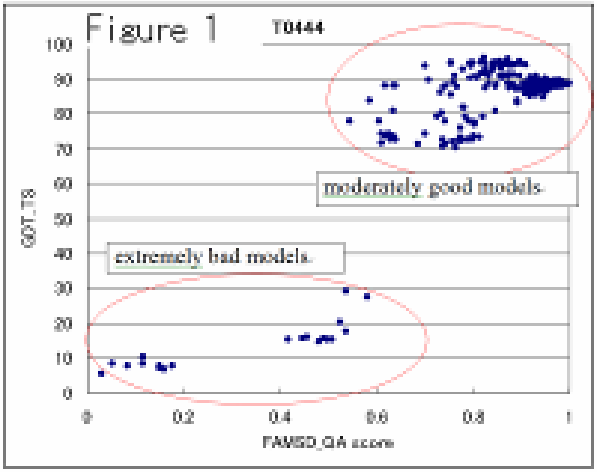
Results and discussion

Correlation coefficients

109 experimental structures of 128 CASP8 targets became available by October 2008. We calculated GDT_TS (accuracy score of backbone geometry) of all predicted models for 103 structure available targets, and calculated Pearson and
Spearman correlation coefficients between GDT_TS and FAMSD_QA score. As a result, average Pearson and Spearman correlation coefficients for all targets were 0.85 and 0.75, respectively. Furthermore the averages for 75 relatively easy targets were 0.91 and 0.79, and for 28 relatively difficult targets were 0.69 and 0.67, respectively. (Target classification is referred to on Robetta evaluation page [5].)

Given this, it can be considered that FAMSD_QA scoring is more effective for easy targets than for difficult targets. The reason for the difference between Pearson and Spearman correlation coefficients for easy targets is that some targets of in the easy category have the bipolar distribution, as shown in Figure 1. Tthere are both moderately good models and extremely bad models. That is, non normal distribution is observed. The target that has the biggest difference between Pearson and Spearman correlation coefficients was T0444, for which the PDB code is 2VUX. These coefficients were 0.857 and 0.289, respectively. Fig. 1 shows the scatter plot of FAMSD_QA score versus GDT_TS. In this case FASMD_QA scoring could judge the moderately good models (GDT_TS > 50) as “good model” and could judge the extremely bad models (GDT_TS < 30) as “bad model”. Therefore Pearson correlation coefficient was very high (0.857). But among the moderately good models, FASMD_QA scoring couldn’t distinguish relatively good models from relatively bad models, so Pearson correlation coefficient calculated with only these models was 0.416 in comparison with 0.857.

This is not so good, but the GDT_TS of the first ranked model by FASMD_QA score is 88.6 and the highest GDT_TS among all models is 96.0. The ratio of the GDT_TS of the first ranked model to the highest GDT_TS, we call MGR (Max GDT_TS Ratio), is 92.3 (88.6/96.0) %. The average MGR value for all targets, easy targets and difficult targets were 89.6, 93.8 and 79.0 %, respectively.



Evaluate the first raked model

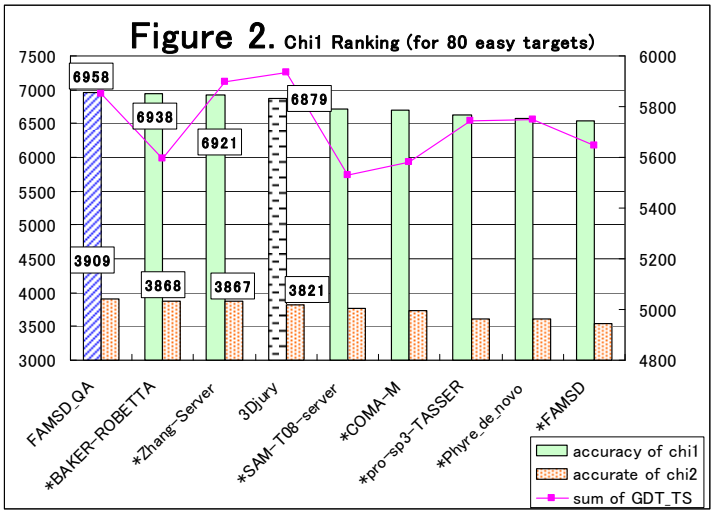
We calculated the cumulative GDT_TS score of the first ranked models by FAMSD_QA score and compared with that of other automatic servers. FAMSD_QA ranked at second following Zhang-Server.

Rank	team name	Sum of GDT_TS	average
1	Zhang-Server	6884.62	63.16
2	**FAMSD_QA**	**6784.26**	**62.24**
3	pro-sp3-TASSER	6654.85	61.05
4	RAPTOR	6614.46	60.68
5	METATASSER	6580.59	60.37

Furthermore, we evaluated the accuracy of side chain torsion angles for easy targets. We calculated the cumulative number of residues that have sufficiently accurate chi1 angle (within 30 degrees from native). As a result, FAMSD_QA ranked first of all server teams and 3D-Jury ranked fourth (Figure 2). This shows that FAMSD_QA scoring can select good models in terms of not only backbone geometry but also side chain torsion angles.

Conclusion

We developed an alternative consensus score for the purpose of selecting good models that have accurate side chain atoms. The new consensus score considers the side chain environment. We participated in Qaulity Assessment category and evaluated our method. As a result, side chain accuracy of the first ranked models by our new method was the best of all servers including 3D-Jury. It was proved that our consensus method using the side chain environment can select better side chain models.



[1] Ginalski K, Elofsson A, Fischer D, Rychlewski L. 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics. 2003; 19:1015–8
[2] Terashi G, Takeda-Shitaka M, Kanou K, Iwadate M, Takaya D, Hosoi A, Ohta K, Umeyama H. Proteins. 2007;69 Suppl 8:98-107.
[3] Ogata, K. and Umeyama, H. An automatic homology modeling method consisting of database searches and simulated annealing J Mol Graph Model. 2000; 18, 258-272
[4] Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 1999; 292: 195-202.
[5] http://robetta.bakerlab.org/CASP8_eval/index.html

Automated homology modeling based upon multiple reference proteins using better pairwise alignments

K. Kanou[1], T. Hirata[1], G. Terashi[1], H. Sakai[1],
M. Takeda-Shitaka[1] and H. Umeyama[1]
School of Pharmacy, Kitasato University, 5-9-1 Shirokane, Minato-ku, Tokyo 108-8641 JAPAN
kanouk@pharm.kitasato-u.ac.jp

We developed an automated method of protein structure prediction called FAMS (Full Automatic Modeling System) [1,2]. FAMS is a homology modeling program consisting of database search and simulated annealing, and can construct high accuracy model when appropriate reference protein was detected. For predicting more accurate model, especially of loop structure and side chain torsion angles, we developed a new version of FAMS, called FAMS-multi, which uses multiple reference proteins. In the following, we describe the scheme of FAMS-multi.

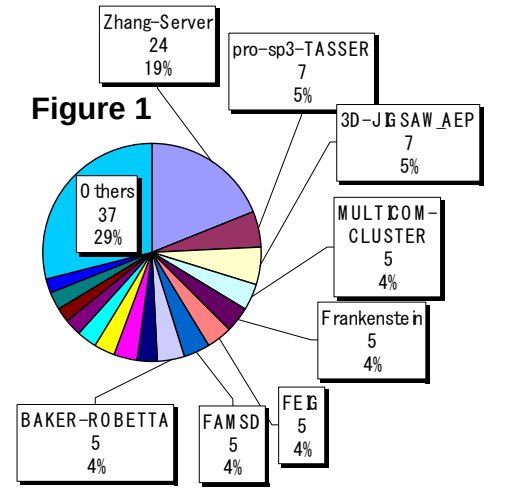


Figure 1

Methods

1. Generation of better pairwise alignments

We used the predicted models by other teams to generate better pairwise alignments between the target and its template in the PDB. First, we rebuilt these models by using FAMS program for the purpose of removing collisions. These rebuilt models were used to generate pairwise sequence alignments between the target and its template. The pairwise alignments were generated by structural superposition between each refined model and the its template using CE program [3]. When the superposition of the model and its template was not performed with the criteria of Z-score > 3.7, the alignment was not used.

Next, we constructed C□ models from these alignments using FAMS-multi program, and calculated 3D-jury scores of these C□ models which is C□□consensus score. Some alignments whose C□ model has a high 3D-jury score were used to construct full atom models using FAMS-multi program, and these models were evaluated using fams-ace2 method. Figure 1 shows the distribution of teams whose alignment was used to construct submitted models.

2. Construction of models by FAMS-multi

Some reference proteins were chosen based on the sequence and structural similarity with template. Next, a multiple structural alignment based on the superposition of C□ atoms was performed among the reference proteins including in the template. The target sequence was put on for this alignment based on the piarwise alignment between target and template mentioned in the preceding section. Thus, we get a result of multiple alignment between a target protein and reference proteins.

Using this alignment, tertiary structures were constructed mainly with next three steps, C□ construction, main chain construction, and side chain construction. In each step, optimization was executed by the simulated annealing method.

C□ construction step: For the initial C□ coordinates, first, the weighted average of C□ coordinates and the average distance were obtained from pairwise structural alignment based on the superposition of C□ atoms of the target and reference proteins. The weight factor of C□ coordinates for each reference proteins was decided based on Local Space Homology (LSH) calculated for each secondary structure segment. Next, the coordinates of C□ atoms were optimized by simulated annealing.

Main chain construction step: Initial coordinates of main chain atoms were constructed with the same method as FAMS. In the simulated annealing step, the potential function, which is consisting of (1) the weighted average of the coordinates of main chain atoms, (2) the average of distance and (3) the pair of N and O atoms forming the hydrogen bond as structural information, was used.

Side chain construction step: For the generated main chain atoms, conserved side chain torsion angles were obtained from homologous proteins. The coordinates of side chain atoms consisting of conserved side chain torsion angles were placed in relation to the fixed main chain atoms. The structural information such as the weighted average of the coordinates, average of distance, and the pair of N and O atoms forming the hydrogen bond, was derived from homologous proteins, and this information was used in optimization procedure.

3. Evaluate models (fams-ace2 method)

Thus, some full atom models were constructed. These models were evaluated using fams-ace2 selecting method (combined C□ □consensus and Circle score [4]). Consequently top five models were selected.

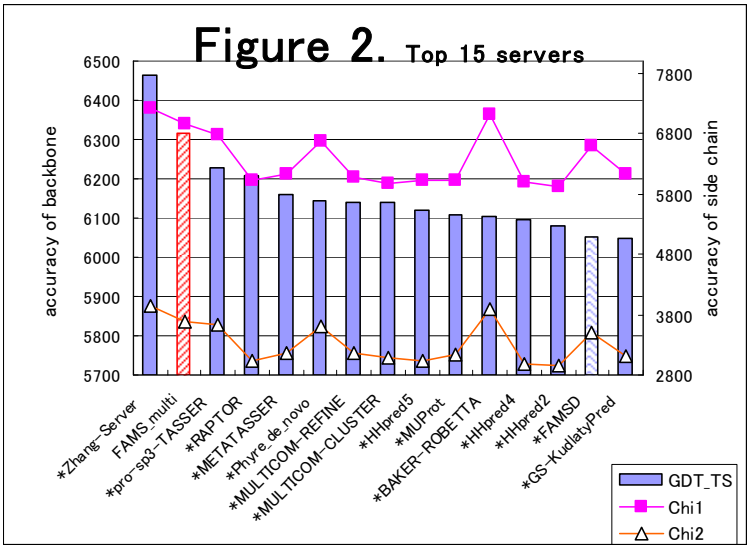
4. Refine models

Five selected models were refined using Energy minimize & Molecular dynamics. With this procedure, hydrogen bonds, main chain torsion angles and side chain torsion angles were refined slightly and collisions of hydrophobic atoms were decreased.

All the procedures were implemented automatically.

Results and discussion

109 experimental structures of 128 CASP8 targets became available by October 3, 2008. We evaluated the accuracy of *FAMS_multi* models and that of the other server models, and compared them. The accuracy of backbone geometry was assessed by GDT_TS score, and the accuracy of side chain was assessed by the number of residues which have a sufficiently accurate side chain (chi1 torsion angle within 30 degrees from native structures or chi2 torsion angle within 60 degrees from native). Figure 2 shows the server ranking with the cumulative GDT_TS score of 109 targets (bar graph). Line graphs of square and triangle point is the cumulative number of accurate Chi1 torsion angles and Chi2 torsion angle, respectively.
As the results, *FAMS_multi* ranked second following Zhang-Server with GDT_TS score. *FAMS_multi* also ranked second following Zhang-Server with side chain accuracy. *FAMS_multi* could construct good models in terms of backbone geometry and side chain conformation.



[1] Ogata, K. and Umeyama, H. J Mol Graph Model 2000; 18, 258-272.
[2] Ogata K, Umeyama H. Proteins. 1998; 31(4):355-69.
[3] Shindyalov IN, Bourne PE. Protein Engineering 1998; 11(9) 739-747.
[4] Terashi G, Takeda-Shitaka M, Kanou K, Iwadate M, Takaya D, Hosoi A, Ohta K, Umeyama H. Proteins. 2007;69 Suppl 8:98-107.

Automated protein structure prediction by comparative modeling and correlation-based scoring

S. M. Gopal[1] and M. Feig[1,2]
[1] *Department of Biochemistry and Molecular Biology,* [2] *Department of Chemistry, Michigan State University, East Lansing, MI; USA*
feig@msu.edu

The most common strategy in protein structure prediction involves two stages:
a) generation of models with some sampling method and (b) evaluation of models with a suitable scoring function to identify the native structure. Often stage (a) is well achieved by numerous well-known methods, whereas design of a robust scoring function has been bottle neck in structure prediction. Decoy scoring with most of the successful scoring potentials (physical/knowledge-based) are noisy due to large fluctuations in physical interactions arising even from minor structural perturbations. Our group has devised a novel correlation-based scoring method[1] which enhances the scoring function by reducing the associated noise. This method takes advantage of the idea that the structure from a given ensemble that is closest to the native basin leads to the highest correlation coefficient between a given score and distance to that structure as an approximation of the native state for the entire ensemble. We apply this scoring method to models generated from diverse comparative, threading and ab initio methods. The following paragraph will briefly summarize the essential steps involved in our automated modeling protocol.

- Alignments were obtained for each domain of given target from different alignment methods such as BLAST[2], FFAS[3], FUGUE[4], HHSEARCH[5], PROSPECTOR[6], SAM[7], SP3[8] and SP4[9]. Each alignment was scored based on sequence identity and predicted secondary structure.
- Models were built from high-scoring single template and multiple template alignments with MODELLER[10] and MMTSB[11] programs. Models were also built from TASSER[12] and ROSETTA[13] ab-initio (only for FM targets) protocols. Further additional models were generated for some of the targets with a local implementation of iterative TASSER method using above models.
- Models were subjected to short minimization and evaluated with DFIRE[14] potential. They were then clustered and clusters with the lower score were used for generating an ensemble of models.
- Decoy scoring in this ensemble was enhanced by reducing the "noise" associated with the score by a correlation-based scoring method[1]. A subset of models with best correlation scores was chosen and the models with lowest DFIRE score from this subset were submitted as predictions.

1. Stumpff-Kane, A. & Feig, M. A Correlation-Based Method for the Enhancement of Scoring Functions on Funnel-Shaped Energy Landscapes (2006). *Proteins,* **63**, 155-164.
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. Jaroszewski, L., Rychlewski, L., Li, Z., Li, W. & Godzik, A. FFAS03: a server for profile-profile sequence alignments (2005). *Nucl. Acids Res.* **33**, W284-W288.
4. Shi, J., Blundell, T. L. & Mizuguchi, K. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties (2001). *J. Mol. Biol*. **310**, 243-257.
5. Söding J. Protein homology detection by HMM-HMM comparison (2005). *Bioinformatics* **21**, 951-960.
6. Skolnick, J., Kihara, D. & Zhang, Y. Development and testing of the PROSPECTOR 3.0 threading algorithm (2004). *Proteins* **56**,502-518.
7. Karplus,K., Karchin,R., Draper,J., Casper,J., Mandel-Gutfreund,Y., Diekhans,M. and Hughey,R. Combining local-structure, fold-recognition, and new fold methods for protein structure prediction (2003). *Proteins*, **53**, 491–496.
8. Zhou, H. & Zhou, Y. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments (2005). *Proteins*. **58**, 321–328.
9. Liu, S., Zhang, C., Liang, S. & Zhou, Y., Fold Recognition by Concurrent Use of Solvent Accessibility and Residue Depth (2007). *Proteins,* **68**, 636-645.
10. Sali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints (1993). *J. Mol. Biol.* **234**, 779-815.
11. Feig, M., Karanicolas, J., & Brooks III, C., L.: MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology (2004). *J. Mol. Graph Model*. **22**, 377-395.
12. Zhang, Y. & Skolnick, J. TASSER: An automated method for the prediction of protein tertiary structures in CASP6 (2005). *Proteins* **61(S7),** 91-98.

13. Rohl, C.A., Strauss, C.E., Misura, K.M.S., & Baker, D. Protein structure prediction using Rosetta (2004). *Methods in Enzymology* **383**, 66-93.
14. Zhou, H. & Zhou, Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction (2002). *Protein Science*, **11**, 2714-2726.

Fiser-M4T

Improved scoring function and template search protocol for comparative modeling using the M4T method

D. Rykunov, E. Steinberger, C.J. Madrid-Aliste and A. Fiser
Department of Systems and Computational Biology, Department of Biochemistry, Albert Einstein College of Medicine, Bronx, NY USA
afiser@aecom.yu.edu

Improvements in comparative protein structure modeling for the remote target-template sequence similarity cases are possible through the optimal combination of multiple template structures and by improving the quality of target-template alignment. Recently developed MMM[1,2] and M4T[3,4] methods were designed to address these problems. MMM identifies alternatively aligned regions from a set of input alignments, maps them in the template structure and scores them using a composite scoring function within the given the structural environment. The final alignment is a combination of the best scored alternatives with the core part of alignment. M4T method implements an algorithm to automatically select and combine Multiple Template structures and feed them to MMM in order to generate a protein model.

Present modification of the MMM method replaces previously used contact statistical potential with recently developed distance-dependent residue-level statistical potential similar to our all-atom Shuffled Reference State potential[5]. It is described in greater details in the **Fiser-QA** abstract. Along with addition of BLOSUM62 mutation table scores it improves alignment accuracy, especially in low sequence identity (<30%) cases.

1 Rai, B.K. and Fiser, A. (2006) Multiple mapping method: A novel approach to the sequence-to-structure alignment problem in comparative protein structure modeling. Proteins: Structure, Function, and Bioinformatics 63, 644-661
2 Rai, B.K. et al. (2006) MMM: a sequence-to-structure alignment protocol. Bioinformatics 22, 2691-2692
3 Fernandez-Fuentes, N. et al. (2007) M4T: a comparative protein structure modeling server. Nucleic Acids Res 35 (Web Server issue), W363-368
4 Fernandez-Fuentes, N. et al. (2007) Comparative protein structure modeling by combining multiple templates and optimizing sequence-to-structure alignments. Bioinformatics 23, 2558-2565
5 Rykunov, D. and Fiser, A. (2007) Effects of amino acid composition, finite size of proteins, and sparse statistics on distance-dependent statistical pair potentials. Proteins: Structure, Function, and Bioinformatics 67, 559-568

Fiser-QA

Assessment of model quality using distance-dependent pairwise statistical potentials with shuffled reference state

D. Rykunov and A. Fiser
Department of Systems and Computational Biology, Department of Biochemistry, Albert Einstein College of Medicine, Bronx, NY USA
afiser@aecom.yu.edu

We developed distance-dependent residue-level statistical potential similar to our all-heavy-atom shuffled reference state (SRS) potential[1]. Key feature of both residue-level and all-heavy-atom potentials is the way system state with no interactions is approximated. Atomic identities for all atom potential and residue identities were shuffled while their spatial positions were preserved. Different sequence separations and distance cutoffs were studied.

This SRS pairwise potential was combined with local potentials developed by Reva and coauthors, including short-range distance-dependent potentials, bend and torsion potentials[2].

All SRS and local potentials were linearly combined with weights adjusted to maximize correlation of score and GDT_TS for comparative model and easy fold recognition targets from previous CASP experiments (CASP5-7). In the quality assessment experiment we have explored three flavors of the potentials:

Fiser-QA server was based on C_β residue-level SRS potential with sequence separation 3 (i.e. pairs *i,i*+1 and *i,i*+2 were excluded) and spatial distance cutoff value 8 Å.

Fiser-QA-Comb server was based on combination of local potentials[2] with C_β residue-level SRS potential (sequence separation 1, spatial distance cutoff value 11 Å).

Fiser-QA-FA potential combined local potentials[2] with all-heavy-atom[1] SRS potential using sequence separation 1, spatial distance cutoff value 4 Å.

All three servers scored models for a given target and normalized scores to 0-1 scale; in addition, Fiser-QA-FA server employed MODELLER[3] to add missing side chain atoms.

1 Rykunov, D. and Fiser, A. (2007) Effects of amino acid composition, finite size of proteins, and sparse statistics on distance-dependent statistical pair potentials. Proteins: Structure, Function, and Bioinformatics 67, 559-568
2 Reva, B.A. et al. (1997) Accurate mean-force pairwise-residue potentials for discrimination of protein folds. Pac.Symp.Biocomput., 373
3 Fiser, A. and Sali, A. (2003) Modeller: generation and refinement of homology-based protein structure models. Methods Enzymol 374, 461-491

fleil

Comparative modeling with all-atom refinement using molecular dynamics simulation

S. Fuchigami[1]
[1] - *International Graduate School of Arts and Sciences, Yokohama City University*
sotaro@tsurumi.yokohama-cu.ac.jp

We have mainly focused on tertiary structure prediction of target proteins categorized into comparative modeling. Our method starts from conventional approaches consisting of template selection, sequence alignment and loop modeling. For some target proteins, we further performed an all-atom refinement of models using energy minimization and molecular dynamics (MD) simulation in explicit solvent.

Template structures for modeling of target sequences were selected by PSI-BLAST[1] searches against the PDB database using position-specific scoring matrices generated by PSI-BLAST with 10 iterations against the nr sequence database. For some targets, we also used information of secondary structure prediction performed by PSIPRED[2] to choose templates. Target sequences were aligned to the templates using PSI-BLAST and with manual curation. Missing loops of target structures were modeled by MODELLER[3].

To remove the atomic clashes in the models, we carried out energy minimization by steepest descents using the MD program system, MARBLE[4], with the CHARMM22 force field for proteins[5] and the CMAP correction for peptide backbone ϕ, ψ dihedral crossterms[6]. Consequently the clashes were considerably reduced to the same extent or less than observed in native crystal structures.

In order to sample possible conformations of the target proteins at atomistic level, we performed MD simulation in NPT ensemble with explicit water, started with the energy-minimized structures, using the MARBLE[4] with the CHARMM22/CMAP force filed parameters[5,6]. The initial structures were dissolved in water molecules with the addition of counter ions to neutralize the net charges of the system. The temperature and pressure of the system were set at 300 K and 1 atom, respectively. Water molecules and hydrogen-containing group (e.g. CH_3, NH_2, OH, etc.) were treated as rigid bodies (partial rigid-body method), enabling to use a 2.0 fs time step. Coulombic interactions were evaluated using the particle-mesh Ewald method[7]. For some targets, additional refinements were carried out using simulated annealing to relax the sampled conformations of the target, especially fluctuating loops

Submitted models were chosen from a set of models generated using different templates and alignments based on complete-linkage clustering, structure verification by WHAT_CHECK[8], and visual inspection.

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. Jones,D.T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292, 195-202.
3. Šali,A. & Blundell,T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234, 779-815.
4. Ikeguchi,M. (2004). Partial Rigid-Body Dynamics in NPT, NPAT and NP□T Ensembles for Proteins and Membranes. J. Comput. Chem. 25, 529-541.
5. MacKerell Jr.,A.D., Brooks,B., Brooks III,C.L., Nilsson,L., Roux,B., Won,Y. & Karplus,M. (1998). CHARMM: The Energy Function and Its Parameterization with an Overview of the Program. in The Encyclopedia of Computational Chemistry edited by Schleyer,P.v.R. et al., (John Wiley & Sons, Chichester, 1998), Vol. 1, pp. 271-277.
6. MacKerell Jr.,A.D. (2004). Empirical Force Fields for Biological Macromolecules: Overview and Issues. J. Comput. Chem. 25, 1584-1606.
7. Essmann,U., Perera,L., Berkowitz,M.L., Darden,T., Lee,H. & Pedersen,L.G. (1995). A smooth particle mesh Ewald method. J. Chem. Phys. 103, 8577-8593.
8. Hooft,R.W.W., Vriend,G., Sander,C. & Abola,E.E. (1996). Errors in protein structures. Nature 381, 272-272.

FLOUDAS

ASTRO-FOLD: Three dimensional structure prediction of proteins using *ab initio* methods

C. A. Floudas[1], A. Subramani[1] , Y. Wei[1] and R. Rajgaria[1]

[1] *Department of Chemical Engineering, Princeton University, Princeton NJ*

floudas@titan.princeton.edu

First principles structure prediction of proteins based on an overall deterministic global optimization framework coupled with mixed-integer optimization. The novel four stage approach uses free energy calculations and integer linear optimization to predict helical and beta-sheet structures. Detailed atomistic modeling and the deterministic global optimization method, aBB, coupled with torsion angle dynamics, form the basis for the final tertiary structure prediction. A hybrid aBB-CSA (conformational space annealing) algorithm has been used to improve performance and generate a large ensemble of low-energy structures[1,2]. The first stage involves the identification of helical segments and is accomplished by partitioning the amino acid sequence into overlapping oligopeptides; atomistic level modeling using ECEPP/3 generates an ensemble of low energy conformations; calculating free energy contributions for each pentapeptide; and helix propensities for each residue using equilibrium occupational probabilities of clusters. A new method for secondary structure prediction is also introduced, wherein an ILP based model for helix prediction has been established, which evaluates the propensity of the central residue of overlapping nonapeptides to be in a helix, by calculating the pairwise probabilities of residues surrounding it[3,4]. The second stage focuses on the prediction of beta-sheet and disulfide bridge topology. It is based on an ILP modeling of the hydrophobic driving force of beta-structure formation and solving the model to maximize the hydrophobic contact energy[5]. In addition, a novel ILP based method for tertiary contact prediction, using a Ca-Ca distance dependent force field to assign contact energy has been implemented[6,7]. The third stage involves the derivation of angle and distance restraints based on helical and beta-sheet predictions to enforce the predicted secondary and tertiary arrangements[8]. Restraints are determined for the loop residues connecting helical and strand regions through dihedral angle sampling and a novel clustering approach[9]. The final tertiary structure prediction relies on restraints introduced from the previous stages as well as atomistic energy modeling, represents a nonconvex constrained global optimization problem, which is solved through the combination of a deterministically based global optimization approach, the aBB, and torsion angle dynamics[10]. A pre-cursor quick rotamer energy optimization step is also implemented[11]. A number of force fields like the High resolution distance dependent force field[12], and the optimized Amber force field[13] were used to re-rank the resulting structures.

1. Klepeis JL and Floudas CA (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.

2. Klepeis JL, Pieja MJ and Floudas CA (2003) Hybrid Global Optimization Algorithms for Protein Structure Prediction: Alternating Hybrids. *Biophys. J.* **84**, 869 – 882.
3. Subramani A and Floudas CA (2008), *in preparation*.
4. Klepeis JL and Floudas CA (2002). Ab-Initio Prediction of Helical Segments in Polypeptides. *J Comput. Chem.*, **23**, 1-22.
5. Klepeis JL and Floudas CA (2003). Prediction of Beta-Sheet Topology and Disulphide Bridges in Polypeptides. *J. Comput. Chem.*, **24**, 191-208.
6. Rajgaria R, McAllister SR and Floudas CA (2008). *in preparation.*
7. McAllister SR, Mickus BE, Klepeis JL and Floudas CA (2006), A Novel Approach for Alpha-Helical Topology Prediction in Globular Proteins: Generation of Interhelical Restraints., *Proteins*, **65**, 930-952.
8. McAllister SR and Floudas CA (2008), Development of rigorous distance bounds for improving protein structure prediction., *submitted*.
9. Monnigmann M and Floudas CA (2005), Protein Loop Structure Prediction with Flexible Stem Geometries. *Proteins*, **65**, 930-952.
10. Klepeis JL and Floudas CA (2003), Ab-Initio Tertiary Structure Prediction of Proteins. *J Glob. Optim.*, **25**, 113-140.
11. McAllister SR and Floudas CA (2008), *in preparation.*
12. Rajgaria R, McAllister SR and Floudas CA (2006). A Novel High Resolution Ca-Ca Distance Dependent Force field Based on a High Quality Decoy Set. *Proteins*, **65(3)**, 726-741.
13. Wroblewska L, Jagielska A and Skolnick J (2008), Development of a Physics-based Force field for the scoring and Refinement of Protein Models., *BioPhys. J.*, **94**, 3227-3240.

GeneSilico

The GeneSilico pipeline for protein structure prediction

J. Orlowski[1]*, M. Boniecki[1], W. Potrzebowski[1], J.M. Bujnicki[1]
1 International Institute of Molecular and Cell Biology in Warsaw, Trojdena 4, 02-109 Warsaw, Poland
jerzyo@genesilico.pl

Based on our experience with the metaserver approach and with modeling by recombination of fragments (e.g. the FRankenstein's Monster protocol), we developed a new prediction method, which makes better use of server models and Model Quality Assessment Programs (MQAPs).

In the first step of modeling, we retrieve all CASP server predictions. The most common template structures are identified. The similarity matrix of the models is calculated using MAXCLUST and models are clustered using CLANS. Models and their fragments are also ranked by their scores predicted by the following MQAPs: PCONS, ModFoldClust, MetaMQAPconsI and MetaMQAPconsII. In the case of MetaMQAPcons methods developed in our laboratory, we use also their special mode of predicting the top 5 models. Simultaneously we carry out the fold recognition analysis through the GeneSilico MetaServer21,which was updated and now includes new methods for prediction of domain boundaries, protein order, protein solvatation and secondary structure.

The second step of our analysis is a manual inspection of the data obtained. At this step, if 1 to 5 starting models are confidently predicted to exhibit a correct fold, these models are subjected to refinement (see below). If no confident fold prediction is made, the target is modeled de novo using Refiner2 and Rosetta 2.33.

In the refinement step, we were looking for poorly scoring or missing protein fragments or fragments with secondary structure different than predicted by the MetaServer. These fragments were refined in up to four ways:

- By copying the coordinates from different models
- By running REFINER with secondary structure and distance restraints
- By running ROSETTA in the loop modeling mode
- By identifying potential errors in the alignment used for generating the parent protein

The final models were selected from variants generated by different modifications according to scores reported by MetaMQAP4 aided by visual inspection.

In order to run ROSETTA automatically, starting from the models generated by the GeneSilico MetaServer, we developed a pipeline of scripts, called ROCKETTA. Briefly, the secondary structure

predictions are taken from the MetaServer, then de novo, loop or loop relax mode is carried out and the resulting decoys are clustered to identify the most promising solutions.

According to the preliminary ranking reported by Zhang and coworkers (http://zhang.bioinformatics.ku.edu/casp8/) our semi-automatic method of protein structure prediction are on average better than all but one automatic predictor. For 6 targets of different difficulty (T0411, T0425, T0437, T0440, T0462_1, T0496_1) our first models were better than any server predictors, which shows that our approach works well for both easy and fold-recognition modeling.

1. Kurowski,M.A., Bujnicki,J.M. (2003). GeneSilico protein structure prediction meta-server. Nucleic Acids Res 31,3305-3307.
2. Boniecki,M., Rotkiewicz,P., Skolnick,J., Kolinski,A. (2003). Protein fragment reconstruction using various modeling techniques. J. Computer Aided Molecular Design 17:725-737
3. Rohl, C. A., Strauss, C. E., Chivian, D., Baker, D. (2004). Modeling structurally variable regions in homologous proteins with rosetta Proteins 55, 656-677.
4. Pawlowski,M., Gajda,M.J., Matlak,R., Bujnicki,J.M. (2008). MetaMQAP: a meta-server for the quality asessment of protein models BMC Bioinformatics 9:403

GRIER-CONSENSUS

Model Validation Using Delaunay Tessellation

John B. Grier
University of North Carolina at Chapel Hill
jgrier@email.unc.edu

The computational geometry technique Delaunay tessellation (DT) was used to determine the fitness of the protein models created by the structure prediction servers participating in CASP8. Delaunay tessellation transforms a set of point in three-dimensional space into an aggregate of space-filling, irregular tetrahedra, known as Delaunay simplices, with the original points being vertices of these tetrahedra. This computational method when applied to protein structures rigorously defines all four nearest neighbor residue clusters within a protein[1]. For CASP8 the coordinates of the side chain centroids of the server models were determined and used as the input of the DT algorithm. This resulted in decomposing all models into four-residue nearest neighbor clusters (or quadruplets). Since each of the four vertices is composed of one of the 20 natural amino acids, each cluster belongs to one of 160,000 different compositional types or motifs. For each of these quadruplets types a weight was given according to its total occurrence in the tessellated protein models. Each model was then scored by summing the weights of the clusters. The score was then normalized with 1.0 being the highest scoring and 0.0 being the lowest scorning model.

1. Zheng, W., Cho, S.J., Vaisman, I.I., & Tropsha, A. (1997). A new approach to protein fold recognition based on Delaunay tessellation of protein structure. Pac. Symp. Biocomput. 486-497.

GS-KudlatyPred

KudlatyPred - Fully automated modeling server based on scoring of models by MetaMQAPcons and recombination of best-scoring fragments

M. Pawlowski and J.M. Bujnicki
Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology in Warsaw, Trojdena 4, PL-02-109 Warsaw, Poland
marcinp@genesilico.pl

For the prediction of the 3D structures in CASP8, we have developed a fully automatic procedure comprising collection of 3rd-party models, local assessment of model quality by MetaMQAPconsI&II developed in our laboratory, and recombination of best-scoring fragments. The whole procedure comprises the following five steps:

1. The method collects models for a given target sequence. By default, the method downloads models based on fold-recognition alignments generated by the GeneSilico metaserver[1]. In the case of FM targets in CASP-8, the method has also downloaded models generated by ROBETTA[2].

2. Each model is scored by MetaMQAPconsI. Both local residue deviations and global model scores are predicted. Five models with the best global score are selected.

3. All input models are divided into partially overlapping fragments containing 1 or 2 secondary structure elements, depending on the target size.

4. All possible combinations of fragments are ranked (without explicitly generating 3D models for each combination). To rank a given combination of fragments, the sum of local MetaMQAPconsI scores is calculated for all residues. In addition, a complex penalty system is applied. Penalty is given for fragments: a) derived from models with different folds; b) derived from models with folds different from the folds of top 5 models selected in step 2; c) if the area of overlap between fragments exhibits different structure.

5. In the last step, 3D models are built for each of 100 top-scored combinations of fragments, using Modeller 9v3[3] in a multi-template mode. Each fragment is considered as a single template with restraints between residues of each fragment and other residues in the initial model from which that fragment was derived. The resulting 100 models are ranked by the MetaMQAPconsII method.

The preliminary results carried out by Zhang and coworkers (http://zhang.bioinformatics.ku.edu/casp8/) suggested that out method is among the 10 top 3D prediction servers according to TM-score and hydrogen bonds score.

1. Kurowski,M.A. & Bujnicki,J.M. (2003). GeneSilico protein structure prediction meta-server. Nucleic Acids Res 31, 3305-7.
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 Suppl 6, 524-33.
3. Sali,A. & Blundell,T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234, 779-815.

GSmetaDisorder

Meta-prediction of intrinsic disorder in proteins

L.P. Kozlowski[1], J. M. Bujnicki[1,2]

[1] - Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology, ul. Trojdena 4, 02-109 Warsaw, Poland, [2] - Laboratory of Bioinformatics, and Biotechnology, Faculty of Biology, ul. Umultowska 89, 61-614, Poznan, Poland

lukaskoz@genesilico.pl

Recent studies have revealed that protein regions lacking 3D structure, so called intrinsic disorder regions, are very common. They are responsible for many protein functions e.g. phosphorylation, transcriptional activation and protein-protein, and protein-DNA binding. The fact that primary structure of ordered and disordered regions is dissimilar (different hydropathy, charge and amino acid composition) enables prediction of disorder/order state directly from the protein sequence. This have been done using statistical methods, artificial neural networks, support vector machines and other approaches. To date, over 20 disorder prediction methods have been published. The best among them achieve less than 80% accuracy.

In order to increase efficiency and robustness of prediction we have benchmarked 13 disorder predictors (DisEMBL, DISOPRED2, DISpro, Globplot, iPDA, IUPred, Pdisorder, Poodle-s, Poodle-l, PrDOS, Spritz, VLS2 and RONN) on two datasets. The first one is the last release of the DisProt database, which stores experimentally validated disorder-containing proteins (470 proteins) and 96 targets from CASP7. The second is culled PDB database employing some restrictions (i.e. length of protein 50-100 amino acids, R-factor < 0.2, resolution < 2.0Å ,and sequence identity < 20%), which gave 1147 proteins. Additionally, other features of sequence are taken into account: amino acid type and its position into sequence, secondary structure prediction (PSIPRED and Jnet) and solvent accessibility (Jnet).

Based on the result of this benchmark we constructed a meta-predictor, which uses as an input predictions from the above-mentioned methods and process them using a backpropagation artificial neuronal network

with single layer. The method uses 9 amino acids long sliding window and smoothing filtering in the final step.

Preliminary results suggest that presented meta-server overperforms each of used methods reaching 83% accuracy.

GS-MetaMQAP
GS-MetaMQAPconsI
GS-MetaMQAPconsII

Model Quallity assessment using MetaMQAP, MetaMQAPcomsI and MetaMQAPconsII

M. Pawlowski and J.M. Bujnicki
Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology in Warsaw, Trojdena 4, PL-02-109 Warsaw, Poland
marcinp@genesilico.pl

Three servers (GS-MetaMQAP, GS-MetaMQAPconsI, and GS-MetaMQAPconsII) from our group participated in QA prediction in CASP8. All these methods predict both local and global model quality.

MetaMQAP[2] evaluates single protein models. MetaMQAPconsI scores models by comparing a series of models with each other. MetaMQAPconsII is a variant of MetaMQAPconsI, which compares all models from the query set against 10 representative models from the query set.

MetaMQAP[2] uses a machine learning approach to assess the deviation of C-alpha atoms of all residues in the model from their counterparts in the unknown native structure. This method combines the output from a number of model quality assessment programs (MQAPs), including VERIFY3D[3], PROSA[4], BALA-SNAPP[5], ANOLEA[6], PROVE[7], TUNE[8], REFINER[9], PROQRES[10], as well as local residue features: secondary structure agreement, solvent accessibility, residue depth. Finally, global deviation is calculated in the form of RMSD and GDT_TS values.

MetaMQAPconsI combines MQAPs scores: VERIFY3D, PROSA, BALA, ANOLEA, PROQRES, DFIRE[11] and local residue features. In addition, input models are superimposed in pairs, then residue contacts and deviation of the C-alpha atoms of corresponding residues are analyzed. Two independent regression predictors were generated. The local predictor inferrs S-scores for each residue in a model, while the globals predictor infers the GDT_TS score of the whole model.

MetaMQAPconsII is a variant of MetaMQAPconsI. Here all models are clustered into 10 groups using the k-mean approach. The clustering threshold is automatically set to a value that guarantees clustering of 75% of initial models. Then, MetaMQAPconsII compares each model to the centroids of these 10 clusters. Thus, MetaMQAPconsII is less sensitive to overrepresentation of similar models in the input dataset.

1. Pawlowski,M., Gajda,M.J., Matlak,R. & Bujnicki,J.M. (2008). Meta-MQAP: a meta-server for the quality assessment of protein models BMC Bioinformatic 9:403.
2. Eisenberg,D., Luthy,R. & Bowie,J.U. (1997). VERIFY3D: assessment of protein models with three-dimensional profiles. Methods Enzymol 277, 396-404.
3. Sippl,M.J. (1993). Recognition of errors in three-dimensional structures of proteins. Proteins 17, 355-62.
4. Krishnamoorthy,B. & Tropsha,A. (2003). Development of a four-body statistical pseudo-potential to discriminate native from non-native protein conformations. Bioinformatics 19, 1540-8.
5. Melo,F. & Feytmans,E. (1998). Assessing protein structures with a non-local atomic interaction energy. J Mol Biol 277, 1141-52.
6. Pontius,J., Richelle,J. & Wodak,S.J. (1996). Deviations from standard atomic volumes as a quality measure for protein crystal structures. J Mol Biol 264, 121-36.
7. Lin,K., May,A.C. & Taylor,W.R. (2002). Threading using neural nEtwork (TUNE): the measure of protein sequence-structure compatibility. Bioinformatics 18, 1350-7.
8. Boniecki,M., Rotkiewicz,P., Skolnick,J. & Kolinski,A. (2003). Protein fragment reconstruction using various modeling techniques. J Comput Aided Mol Des 17, 725-38.
9. Wallner,B. & Elofsson,A. (2006). Identification of correct regions in protein models using structural, alignment, and consensus information. Protein Sci 15, 900-13.

10. Zhou,H. & Zhou,Y. (2002). Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 11, 2714-26.

Hamilton-Torda-Huber

Protein Contact Prediction Using Patterns of Correlation

N.A. Hamilton[1], A.E. Torda[2] and T. Huber[3]
[1]– Institute for Molecular Bioscience, The University of Queensland,
[2]– Zentrum für Bioinformatik, Universität Hamburg,
[3] – School of Molecular and Microbial Sciences, The University of Queensland
n.hamilton@imb.uq.edu.au - torda@zbh.uni-hamburg.de - t.huber@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 information that can be used, for instance, to rank models created by other methods. To assess progress made in contact predictions in CASP8 predictions we have used identical methods and databases as in CASP7. In the following we briefly describe our 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
The 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.

Handl-Lovell

De novo prediction using multiobjective iterated local search

J. Handl and S.C. Lovell[2]
[1] – *University of Manchester*
simon.lovell@manchester.ac.uk

We describe a novel optimization approach to de novo protein structure prediction that combines iterated local search and multi-objective optimization.

One of the fundamental bottle-necks in de novo protein structure prediction is conformational sampling: due to the high-dimensional and multimodal nature of the landscapes described by current energy functions, even powerful search techniques often converge to local optima and fail in identifying the minimum energy structures. One popular approach taken to ameliorate this problem is the generation of a large number of decoys, each obtained through a random restart of the search technique used. While this usage of random restarts has been shown to be successful in improving the overall search outcome, the approach is limited in the sense that restarts are performed independently of each other and that the information obtained by previous restarts is not exploited. In our work, we have designed an algorithm that maintains an archive of the best structures generated so far, and uses this knowledge to guide future search.

Our method makes use of the principles of iterated local search. Iterated local search is a powerful meta-heuristic, which is based on two key mechanisms: (i) a local search heuristic is used to locally optimize a given solution; and (ii) the optimized solution is perturbed in order to escape local optima; the perturbed solution is then further optimized using the local search heuristic. The search heuristic used in our approach is the Rosetta[2] software for protein structure prediction, and Rosetta can be used to achieve both of the above effects, i.e. to perform random restarts and to achieve the perturbation and further optimization of a given solution. Specifically, in a given iteration, our algorithm chooses between two different types of steps: (i) with probability p, it performs a random restart of Rosetta; or (ii) with probability 1-p, it uses Rosetta to resume search near one of the solutions currently in the archive. The probability p decreases as the algorithm progresses.

We find that the quality of a given structure is more reliably assessed using a range of criteria (rather than the Rosetta low resolution energy only), and our algorithm therefore employs a multiobjective formulation. Specifically, candidate structures are evaluated with regard to four objective functions, namely the Rosetta low resolution energy, a short-range and a long-range hydrogen term and the radius of gyration (as implemented in Rosetta). A candidate structure A is said to be dominated by another structure B if A is worse or equal than B under all objectives, and if A is worse under at least one objective. At a given moment in time, the archive of our method only maintains those solutions that are not dominated by any other candidate structure generated so far. At the end of the run, the algorithm returns the set of structures in the final archive (rather than a single solution, as done by traditional scalar optimization techniques).

Experimental results have shown that the algorithm successfully generates decoy structures that are significantly lower in energy (and also significantly better with respect to the other three objectives) than those obtained from standard runs of the Rosetta method. In the CASP experiment, our multiobjective iterated local search was used for predictions in the free modeling category. For each prediction, the

algorithm was run 5 times for 5000 iterations, corresponding to 25000 calls of Rosetta. The final archives from all five runs were combined and filtered to exclude dominated solutions. The Model 1 to Model 5 submissions were selected from this set using a fully automated technique that considered solutions that performed particularly well under specific pairs of objectives.

1. Simons, K.T., Kooperberg C., Huang, E., Baker, D. (1997) Assembly of protein tertiary structures from fragments with similar local sequences. *J. Mol. Biol.*. **268**, 209-25.

IBT_LT

Template-based modeling of CASP8 target proteins using automatic tools and human expertise

Č. Venclovas and M. Margelevičius
Institute of Biotechnology, Graičiūno 8, Vilnius, Lithuania
venclovas@ibt.lt

In human expert mode we used results of our two automatic servers COMA and COMA-M as a starting point. Both servers are based on a newly developed profile-profile search and comparison method (manuscript in preparation). The difference between the two is that for model-building COMA-M can combine multiple templates, while COMA is taking just a single best template. The servers are described in more detail in a separate abstract.

The degree of human intervention varied depending on the initial assessment of models produced by COMA/COMA-M in the context of results by other automatic methods participating in CASP8. If models had no obvious flaws and fared well relative to those obtained by other automatic methods, little or no human intervention was used. However, if the model assessment was suggesting that COMA/COMA-M models could be further optimized, both template selection and/or alignment were manually refined using additional techniques as described below.

Template selection
Template selection was usually based on the consensus results of transitive PSI-BLAST[12] searches using the PSI-BLAST-ISS tool[13]. In other words, the most representative structure of the target sequence family was considered to be the best template. If there were multiple templates, several structures that introduce sufficient conformational variability were selected. In high sequence similarity cases only BLAST results were used to choose template(s). When PSI-BLAST-ISS failed to detect any templates, those detected by COMA and other automatic servers were used. In cases of multiple templates, the selection was an iterative process and the final set of templates was dependent on evaluation of corresponding models.

Sequence-structure alignments
Unless the alignment was trivial, reliably aligned regions were first identified with PSI-BLAST-ISS[13]. In parallel, automatic server models were aligned with one of the representative templates using DaliLite[14] and all the corresponding pairwise alignments were merged into a single, PSI-BLAST-ISS-like alignment. Again, a good agreement between different models was considered to be an indicator of a reliable sequence-structure alignment region. For the remaining (unreliably aligned) regions alternative alignment variants were evaluated at the level of 3D models and the best variant was used in a final model.

Model construction
Three-dimensional structures were constructed automatically from sequence-template(s) alignments using MODELLER[15]. Residue side chains were positioned with SCWRL3[16]. No further optimization was performed.

Model evaluation
Models based on alternative template sets and/or alternative alignments were assessed by several methods including Prosa2003[4] profiles and Z-scores, and visual inspection. Optimization of both the set of templates and the sequence-structure alignment was performed in an iterative manner until model scores could not be improved anymore and a final model looked acceptable by visual analysis.

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. Margelevičius, M. & Venclovas, Č. (2005). PSI-BLAST-ISS: an intermediate sequence search tool for estimation of the position-specific alignment reliability. BMC Bioinformatics 6, 185.
3. Holm, L. & Park, J. (2000). DaliLite workbench for protein structure comparison. Bioinformatics 16, 566-7.
4. Šali, A. & Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234, 779-815.
5. 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.
6. Sippl, M. J. (1993). Recognition of errors in three-dimensional structures of proteins. Proteins 17, 355-62.

Infobiotics

Reside-residue contact prediction using a large-scale ensemble of rule sets and the fusion of multiple predicted structural aspects

J. Bacardit[1,2], M. Stout[1,2] , P.Widera[1], J.D. Hirst[3] and N. Krasnogor[1]
[1] - School of Computer Science, University of Nottingham, [2] – School of Biosciences, University of Nottingham, [3] – School of Chemistry, University of Nottingham
natalio.krasnogor@nottingham.ac.uk

We have constructed an ensemble of more than one thousand rule sets to participate in the residue-residue contact category of CASP. The rule sets were generated by our in-house machine learning system, BioHEL[1]. Three types of input information were used: (1) detailed information of three windows of residues centered at specific points within the protein sequence. (2) information about the connecting segment between the two target residues and (3) global sequence information.

There are two windows of ±4 residues around the two target residues and one window of ±2 residues around the middle point in the chain between the two target residues[4]. For each residue in all of the three windows we included (1) a position-specific scoring matrix (PSSM) profile, computed with PSI-Blast[3], (2) predicted secondary structure using PSIPRED[5], (3) predicted five-state coordination number (CN)[6], (4) predicted five-state relative solvent accessibility (SA)[1] and (5) predicted five-state Recursive Convex Hull (RCH)[1]. CN, SA and RCH were predicted using BioHEL.

The connecting segment is represented by the distribution of amino acids and predicted secondary structure states[4], as well as the distributions of predicted CN, SA and RCH. The global sequence information included the sequence length and the distributions, for the whole sequence, of amino acids and predicted SS, SA, RCH and CN, the number of residues of separation between the two target residues[4] and the contact propensity between the amino acid types of the two target residues[7]. In total 631 variables were used for the training process. The training process followed the following steps:

1. We selected a set of 2811 protein chains from PDB-REPRDB with a resolution less than 2Å, less than 30% sequence identify and without chain breaks nor non-standard residues. 90% of the proteins (~490000 residues) were used for training, 10% for test. BioHEL was trained to predict RCH, SA and CN using the same 90% of proteins
2. For the residue-residue prediction, this 90% was randomly halved, additionally removing any chain longer than 350 residues. Still, the training set contained 15.2 million pairs of residues, from which less than 2% were real contacts
3. To create balanced training sets (in terms of contacts/non contacts) we randomly sampled the these 15.2 million pairs 50 times to create 50 training sets, where each set contained around 300000 residue pairs with a fixed 2:1 proportion of non-contacts to real contacts.
4. We run BioHEL 25 times for each training sample with different initial random seeds, thus generating an ensemble of 1250 rule sets (50 training samples x 25 seeds) to perform the residue-residue contact prediction.

1. Stout, M., Bacardit, J., Hirst, D. and Krasnogor, N.. (2008) Prediction of Recursive Convex Hull Assignments for Protein Residues. Bioinformatics 24(7):916-923.
2. Noguchi, T., Matsuda, H., and Akiyama, Y.. (2001). Pdb-reprdb: a database of representative protein chains from the protein data bank (pdb). *Nucleic Acids Res*, 29:219–220.

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, 3389-402.
4. Punta, M. and Rost, B. (2005) "Profcon: novel prediction of long-range contacts". Bioinformatics 21(13):2960-8.
5. Jones, D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* ***292***, 195-202.
6. Bacardit, J., Stout, M., Hirst, J.D., Krasnogor, N. and Blazewicz, J.. (2006) Coordination Number Prediction using Learning Classifier Systems: Performance and Interpretability. Proceedings of the 8th Annual Conference on Genetic and Evolutionary Computation (GECCO2006), pp. 247-254, ACM Press
7. Shackelford, G. and Karplus, K.. (2007) Contact Prediction using Mutual Information and Neural Nets. *Proteins: Structure, Function, and Bioinformatics*, 69(S8):159-164.

Jones-UCL

FRAGFOLD and BioSerf: developing methods for manual and automatic prediction of novel protein folds

M.I. Sadowski, S. Ward, K. Bryson and D.T. Jones
Bioinformatics Group, Department of Computer Science, University College London, Gower St., London, WC1E 6BT, United Kingdom
d.jones@cs.ucl.ac.uk, URL: http://bioinf.cs.ucl.ac.uk

The Jones-UCL group's main efforts in CASP8 were in improvements to our fragment assembly method (FRAGFOLD[1]) and a new fully automatic server for protein structure structure prediction and modelling (BioSerf). In addition we ran a number of existing servers for disorder prediction (DISOPRED2 [2]), fold recognition (mGenTHREADER [3]), domain prediction (DomPred [4]) and model quality assessment (MODCHECK-HD [5] and MODCHECK-Jury).

For CASP8 target domains which we believed could not be reliably predicted using fold recognition methods, FRAGFOLD 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. FRAGFOLD4 differs from previous versions mainly in the areas of improved long-range hydrogen-bonding and improved fragment selection, including a broader range of fragment types and lengths than in previous implementations. A new option to constrain distances within a model has also been implemented and used for some multidomain targets. As many as 5000 structures were generated for each target domain using a 300 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.

The mGenTHREADER method has been significantly improved since CASP7 (paper in preparation), but there are no significant changes in our DomPred and DISOPRED2 servers from the last CASP experiment. The two model quality servers are new. The MODCHECK-HD server is intended to rank close to native models and is based on a range of detailed atomic preference scores, whereas the MODCHECK-Jury method is aimed at selecting models according to the degree of structural clustering. MODCHECK-Jury makes use of a very simple superposition jury score and also produces residue-by-residue accuracy estimates.

BioSerf, our new fully-automatic protein structure prediction server, was designed to perform structure predictions for a protein with any level of homology with known structures. As such, it uses a pipeline approach, with multiple potential template candidates identified and aligned using first BLAST, then PSI-BLAST[6]. Secondary structure was predicted using PSIPRED[7] , and lastly more template candidates were collected and aligned via mGenThreader. For each of those steps, trivial homologues with full coverage of the target terminated the template search, and a model was created using MODELLER[8] . In cases where full coverage of the protein was not possible from reasonable confidence template predictions, including using multiple template alignments, the remaining fragments (and in some cases, complete proteins split in chunks of up to 150 residues) were used as input to FRAGFOLD. Additionally, the group of closest homologues were used to predict potential contacts so as to provide an additional energy model term to further improve the quality of the ab initio models. These predicted fragments were then clustered, with the representative closest to the largest centroid being used as an additional aligned template for the final

model prediction. In cases where the entire structure had no suitable templates, multiple candidate structures were returned, based on the centroids for the top two clusters, as well as the top scoring models for several energy terms.

1. Jones D.T. (1997) Successful ab initio prediction of the tertiary structure of NK-Lysin using multiple sequences and recognized supersecondary structural motifs. PROTEINS. Suppl. 1, 185-191.
2. 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.
3. McGuffin, L.J. & Jones, D.T. (2003) Improvement of the GenTHREADER method for genomic fold recognition. Bioinformatics, 19, 874-881.
4. Bryson, K., Cozzetto, D. & Jones, D.T. (2007) Computer-assisted protein domain boundary prediction using the DomPred server. Curr Protein Pept Sci., 8, 181-188.
5. Sadowski, M.I. & Jones, D.T. (2007) Benchmarking template selection and model quality assessment for high-resolution comparative modeling. Proteins. 69, 476-485.
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. Jones, D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292, 195-202.
8. A. Sali & T.L. Blundell. (1993) Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234, 779-815.

KOLINSKI

3D Structure Prediction Using a Combination of 3D Lattice Threading de novo Modeling and All-atom Structure Refinement

A. Kolinski[1], A. Zwolinska[1], M. Jamroz[1], A. Rutkowska[1], S. Trojanowski[1] and P. Pokarowski[2]

[1]*Laboratory of Theory of Biopolymers, Faculty of Chemistry, University of Warsaw, Poland*
[2] *Institute of Informatics, Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Poland*
kolinski@chem.uw.edu.pl

In the first stage all targets were divided into three groups: these with dependable structural templates, these with poor and/or fragmentary templates and these for which we decided not to use any templates. The classification was based on the results from metaservers and on our analysis of structural consistency of the emerging templates. The three groups of targets were processed separately using different modeling pipelines. In all cases variants of newly developed CABS-based high resolution lattice modeling tools were employed.

For the CM targets large sets of distance restraints were extracted from many (when available) templates and weighted by a newly developed procure based on local and global similarities between templates. Next Replica Exchange Monte Carlo (REMC) simulations were performed using the CABS engine[1]. The starting replicas were built basing on the templates' structures. The resulting trajectories were clustered[2] and the clusters' centroids were subject to energy minimization[3] after the all-atom reconstruction of the lattice structures[4]. Finally, five models were selected basing on the cluster sizes and the energy of the energy-minimized structured. Thus, this pipeline was somewhat similar to that employed by Kolinski-Bujnicki group during CASP6[2].

The border-line targets, with poor templates were processed using the new CABS-based modeling tool[5] called TRACER. TRACER uses a single template. Projected onto the CABS lattice the template provides a multi-featured three-dimensional scaffold, where various local properties of the template structure are assigned to the lattice points, including amino acid identities, chain direction, secondary structure assignment and hydophobicity profiles. The target chain is allowed to move around the space occupied by the template using the CABS force field and the local comparison of template properties and properties of the target chain. Thus the TRACER procedure is a unification of a true three dimensional threading with de novo modeling and does not require an initial alignment. An optimal (usually fragmentary) alignment of the emerging target structure with the template is successively build during the REMC simulations. Fold selection and refinement is done in a similar fashion as it was done for the CM targets.

The template-free de novo modeling also employs the CABS tools. The simulations were done either without any distance restraints or with a small number of weak restraints imposed onto strongly predicted elements of secondary or supersecondary elements. In the cases of the template-free modeling larger sets of clusters from REMC simulations were analyzed and scored in order to detect the most-likely target structures. We would like to note that CABS free modeling allows also for dependable folding mechanism predictions for small proteins[6-7].

In all cases we did not perform any manual alignment adjustments or the expert-based structure corrections. The procedures are essentially automated.

1. Kolinski,A. (2004). Protein modeling and structure prediction with a reduced representation. Acta Biochimica Polonica 51, 349-371.
2. Gront,D. & Kolinski,A. (2005). HCPM – program for hierarchical clustering of protein models. Bioinformatics, 21, 3179-3180.
3. Kmiecik,S., Gront,D & Kolinski,A. (2007). Towards high-resolution structure prediction. Fast refinement of reduced models with all-atom force field. BMC Structural Biology 7, 43.
4. Gront,D., Kmiecik,S. & Kolinski,A. (2007). Backbone building from quadrilaterals. A fast and accurate algorithm for protein backbone reconstruction from alpha carbon coordinates. J. Comput. Chem. 28, 1593-1597.
5. Kolinski,A. & Gront,D. (2007). Comparative modeling without implicit sequence alignments. Bioinformatics 23, 2522-2527.
6. Kmiecik, S. & Kolinski,A. (2008). Folding pathway of the B1 domain of protein G explored by a multiscale modeling. Biophys. J. 94, 726-736.
7. Kmiecik, S. & Kolinski,A. (2007). Characterization of protein folding pathways by reduced-space modeling. Proc. Natl. Acad. Sci. USA 104,12330-12335.

LCBContacts

Predicting residue-residue contacts using hidden Markov models trained on local neighborhoods of protein structure

P. Björkholm[1,5], P. Daniluk[2], A.Kryshtafovych[3], K. Fidelis[3], R. Andersson[1] and T.R. Hvidsten[1,4]

[1] - The Linnaeus Centre for Bioinformatics, Uppsala University, Uppsala, Sweden, [2] - Department of Biophysics, Faculty of Physics, University of Warsaw, Warsaw, Poland, [3] - UC Davis Genome Centre, UC Davis, USA, [4] - Umeå Plant Science Centre, Department of Plant Physiology, Umeå University, Umeå, Sweden, [5] - Stockholm Bioinformatics Center, Stockholm University, Stockholm, Sweden

torgeir.hvidsten@plantphys.umu.se

Correct prediction of residue-residue contacts in template-free targets would bring *ab initio* protein structure prediction a large step forward. The lack of such correct contacts, and in particular long-range contacts, is considered the main reason why these methods fail[1]. Thus residue-residue contact prediction is an important bioinformatics research area[2] that could help identify the structures that are not reachable by homology modeling.

We propose a novel hidden Markov model based method for predicting residue-residue contacts from protein sequences that is trained on homologous sequences, predicted secondary structure and a library of local neighborhoods (local descriptors of protein structure)[3]. The structural neighborhoods are composed of sets of at least three backbone fragments that are in proximity to each other in space but not necessarily along the amino acid sequence. These structural entities thus incorporate short-, medium- and long-range contacts between different backbone fragments. We used a library of 7151 commonly recurring local descriptors (local descriptor groups) general enough to allow reassembly of the cores of nearly all proteins in the PDB.

HMMs are used to model local descriptor groups. Each position in the multiple alignment of structurally matching descriptors is modeled as a match state while the rest of the sequence (not matching the local descriptor) is modeled by insert states. Some groups may contain fragments of varying length because only parts of the fragments structurally match the group according to the defined similarity threshold[3]. This is handled by using delete states that are tied to specific match states. In order to ensure that whole fragments are not deleted there are two different types of delete states that are disconnected; delete states that are located in the beginning of the fragments and delete states that are located at the end of the fragments. We do not expect a significant sequence signal from these structurally unmatched positions, implying that the delete states have the same emission probabilities as the insert states.

The Viterbi algorithm was used to obtain the most probable alignment between local descriptor groups and a target sequence. We found that the best approach to discriminate targets that contain a local descriptor and targets that do not was to consider the sum of the log values from the match and delete state emissions/transitions only. This eliminates the problem of accounting for different sequence lengths when comparing scores from different targets. Each HMM was matched to the target and accepted if the Viterbi score was higher than an associated threshold shown to discriminate relevant targets in the training set. Contacts were then transferred to the target from the corresponding local descriptor group recognized by the HMM. Only contacts between residues located in different backbone fragments were considered. Each predicted contact was then given a score equal to the sum of the scores from all HMMs predicting that contact. Thus, contacts predicted by many different local descriptor groups were given a higher score than contacts predicted by fewer models. For each type of contact (short-, medium- and long-range) we chose the $N_{pred} = Pct \cdot L$ best predictions, where L was the sequence length and Pct equaled 0.2, 0.5 or was taken from a spline fitting the actual distributions of contacts in proteins with known structure.

1. Floudas,C.A., Fung,H.K., McAllistera,S.R., Mönnigmanna,M. & Rajgariaa,R. (2006) Advances in protein structure prediction and de novo protein design: A review Chemical Engineering Science. 61, 966-988.
2. Sitao,W., & Yang,Z. (2008) A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. Bioinformatics, 24, 924-931
3. Hvidsten,T. R., Kryshtafovych,A., & Fidelis,K. (2008) Local Descriptors of protein Structure: A systematical analysis of the sequence-structure relationship in proteins using short- and long-range interactions. Proteins: Structure, Function, and Bioinformatics, In Press.

LEE

Protein Structure Prediction based on Global Optimization and Alternative Alignment assisted by Quality Assessment

K. Joo[1], J. Lee[1,2], M. Oh[1], S. Shin[1], J. Ko[3], D. Lee[3], H. Park[3], I. H. Lee[1,4], C. Seok[1,3], S. J. Lee[5] and Jo. Lee[1*]

[1]*School of Computational Sciences, Korea Institute for Advanced Study, Seoul 130-722, Korea*
[2]*Department of Mathematics, Kwangwoon University, Seoul 139-701, Korea*
[3]*Department of Chemistry, Seoul National University, 151-747, Korea*
[4]*Korea Research Institute of Standards and Science (KRISS), 305-600, Korea*
[5]*Department of Physics, Suwon University, 445-743,Korea*
* jlee@kias.re.kr

In order to predict the three dimensional structures of all 127 CASP8 targets, we have developed a method that is based on global optimization of score functions in three stages of modeling[1]. For a given set of templates, we have considered many alternative alignments assisted by quality assessment. The whole procedure is composed of the following 7 steps:

1. Fold recognition: To collect fold candidates of a given target sequence, we considered top scoring templates from the meta-server provided by http://bioinfo.pl/~3djury, as well as from the in-house fold recognition method called FoldFinder. FoldFinder is a profile-profile alignment method utilizing predicted secondary structures. We have used a fold database of 17163 protein chains obtained from PISCES[2] at the 95% sequence identity level. With these templates (we have considered up to 25 top-scoring templates), we performed a structural clustering from which typically 5 to 10 sets (lists) of templates are generated. These lists are the input to the following procedure.

2. Multiple sequence/structure alignment by MSACSA[3]: We perform multiple sequence/structure alignment for each template list obtained from the fold recognition step. Unlike the other heuristic (progressive) alignment methods popular in the literature, we have applied a more thorough global optimization method to an in-house consistency-based scoring function similar to the COFFEE by using the conformational space annealing[4] (CSA) method. We have constructed a pair-wise restraint library generated from profile-profile alignment between the query sequence and template sequences and structure-structure alignment between templates using TM-align[5]. Typically, a total of 100 alignments are generated for each list.

3. First-level quality assessment to screen high-scoring alignments: For each alignment, we generate 25 protein 3D models using MODELLER[6] and these models are used to measure the quality of the alignment. For the quality assessment, we trained a support vector regression machine using feature vectors composed

of MODELLER energy, DFIRE energy, secondary structure propensity, solvent accessibility, hydrophobicity and in-house implementation of selected TASSER energy terms. For training, decoy structures generated from our CASP7 models are used. Typically, a total of 10-20 alternative alignments are chosen from the whole lists.

4. 3D structure modeling by MODELLERCSA[7] using the alternative alignments: 100 3D structures of a target protein are generated for each alignment by optimizing the MODELLER energy function using the CSA method. The resulting 1000-2000 models are input for the second-level quality assessment in the following step.

5. Second-level quality assessment and structure clustering to select the best 5 models: We perform the identical quality assessment used in the step 3 to select the best alignment which produces best models on average according to the quality assessment score. The best 5 models are selected by structure clustering of the 100 models from the winning alignment.

6. Re-modeling of insertion regions by a loop modeling method that utilizes fragment assembly and analytical loop closure: Insertions were identified from the sequence alignment and modeled so that the loop closure constraint is satisfied and deviation from fragment structures is minimized simultaneously. DFIRE score was used to select the best loop model.

7. Side-chain modeling of the five selected models by ROTCSA: For side-chain modeling of each selected model, first, a rotamer library is constructed based on the consistency of the side chains from the final 100 models obtained in the step 4. Into this library, we add a backbone dependent and sequence specific rotamer library similar to the SCWRL3.0[8]. Finally, using the CSA, we optimize an in-house scoring function which contains energy terms from SCWRL and DFIRE.

1. Joo,K., Lee,J., Lee,S., Seo,J., Lee,S.J. & Lee,J. (2007) High-accuracy template based modeling by global optimization. *Proteins*, **69**(S8), 83-89.
2. Wang,G. & Dunbrack, Jr. R.L. (2003) PISCES: a protein sequence culling server. *Bioinformatics*, **19**, 1589-1591.
3. Joo,K., Lee,J., Kim,I., Lee,S.J. & Lee,J. (2008) Multiple sequence alignment by conformational space annealing. *Biophys J.*, published online.
4. Lee,J., Scheraga,H.A., & Rackovsky,S. (1997) New optimization method for conformational energy calculations on polypeptides: Conformational space annealing. *J. Comput. Chem.* **18**, 1222-1232.
5. Zhang,Y., & Skolnick,J. (2005) TM-align: A protein structure alignment algorithm based on TM-score. *Nucleic Acids Res.* **33,** 2302-2309.
6. Sali,A. & Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol*. **234**, 779-815.
7. Joo,K., Lee,J., Seo,J., Lee,K., Kim,B. & Lee,J. All-Atom Chain-Building by Optimizing MODELLER Energy Function Using Conformational Space Annealing. *Submitted.*
8. Canutescu,A.A., Shelenkov,A.A. & Dunbrack,Jr. R.L. (2003) A graph theory algorithm for protein side-chain prediction. *Protein Science* **12**, 2001-2014.

LevittGroup

Prediction pipelines for refinement, HM and ab-initio targets

N. Kalisman, A. Shmygelska, G. Chopra, S.Moreno and M. Levitt
Dept. of Structural Biology, School of Medicine, Stanford University

In CASP8 we submitted predictions for all the targets by processing them through three different pipelines. The Template-based Pipeline processed all the targets with PSI-BLAST[1] E-value below 0.01 or 3D-Jury[2] score above 45. The Model-selection Pipeline processed all the other targets. In addition, a separate technique was tested on the refinement targets. Nearly all the steps in the pipelines are amenable to automation, and we hope to run them under the server category at future CASPs.

Template-based Pipeline – The top scoring PSI-BLAST or 3D-jury hit was used as our single template. We performed minor manual adjustments on the raw alignments for most targets. Regrettably, these adjustments eventually proved to be deteriorating with two exceptions: 1) A correction of an obvious alignment error in a membrane protein targets. 2) Occasional elongation of the alignment on either side that was missed due to the local nature of PSI-BLAST. We modeled de novo the gaps in the alignment, as well as regions that had very low similarity to the template with our novel loop-building protocol. This protocol assembles loops from fragments chosen according to backbone conformation propensity[3]. The resulting

ensemble of loops was clustered and scored with the MESHI[4] force field. The lowest energy representatives were incorporated into the final models.

Model-selection Pipeline – We derived a composite and optimized energy function for the ranking of models in cases where a template to the target sequence cannot be reliably determined. The new energy function ranked the models submitted by the participating servers. Selected models were further refined by minimization under the knowledge-based potential of ENCAD[5] prior to submission.

Utilizing one of the most efficient machine learning approaches that prevents over-fitting the data – Support Vector Regression (SVR), we derived and tested a set of weights for a number of established energy functions and their individual terms such that the resulting composite energy function correlates best with the GDT_TS scores of the models. We used both low- and high-resolution knowledge-based and molecular dynamic atomistic potentials from ENCAD[5], Rosetta[6] and MESHI[4]. We trained and tested our SVR using cross-validation on a representative data set from CASP7 targets (112 targets, 16410 server models). We found that a major part of the discriminatory power of the learned composite energy function comes from the following four energy terms: Rosetta environment score, MESHI full solvation score, Rosetta de novo low-resolution centroid score, and ENCAD knowledge-based score.

Model Refienement – We tested direct energy minimization under various potentials as a mean for further refining the proposed models. The most promising potential was KB012, the ENCAD differentiable knowledge-based potential, which was used to generate the #1 submitted models. KB01 is a mean-field approach for the atomic-pair interactions in proteins. Training on the CASP7 refinement category led to improved GDT_TS score on all targets.

1. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W & Lipman,DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389-3402.
2. Ginalski K, Elofsson A, Fischer D, and Rychlewski L (2003) 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics. 19:1015-1018.
3. Shortle D. (2003) Propensities, probabilities, and the Boltzmann hypothesis. Protein Sci. 12:1298-1302.
4. Kalisman, N., Levi, A., Maximova, T., Reshef, D., Zafriri-Lynn, S., Gleyzer, Y., Keasar, C. (2005) MESHI: a New Library of Java Classes for Molecular Modeling. Bioinformatics. 21:3931-3932.
5. Summa, C. M., Levitt, M., (2007) Near-native Structure Refinement Using in Vacuo Energy Minimization. Proc. Natl. Acad. Sci. U.S.A. 104:3177-3182.
6. Rohl CA, Strauss CE, Misura KM, Baker D. (2004) Protein structure prediction using Rosetta. Methods Enzymol. 383:66-93.

LOOPP

LOOPP: A server for sensitive detection of structural templates and homology modeling

B. K. Vallat[1], J. Pillardy[2], T. Blom[1], P. Majek[1,3], B. Cao[1], J. Meller[4,5] and R. Elber[1]

[1] Institute for Computational Engineering and Sciences, University of Texas at Austin, Austin, TX

[2] Computational Biology Service Unit, Core Laboratories Center and Center for Advanced Computing, Cornell University, Ithaca, New York 14853

[3] Department of Computer Science, Cornell University, Upson Hall 4130, Ithaca, New York 14853

[4] Division of Biomedical Informatics, Children's Hospital Research Foundation, Cincinnati, Ohio 45229

[5] Department of Informatics, Nicholas Copernicus University, 87-100 Toru, Poland

brindakv@gmail.com

LOOPP is a homology modeling server. It is based on a template detection algorithm learned by mathematical programming techniques that combines a large number of signals and significantly enhances typical detection capabilities (PSI-BLAST) by about 50 percent. It also uses a novel algorithm for alignment, and it finally builds atomically detailed models with Modeller (using the identified templates and our alignments of the target sequence into them). The strength of the algorithm is (perhaps) in the very large training and test sets that we developed and use. Weaknesses include misses of some trivial PSI-BLAST signals (the training emphasizes difficult signals). The algorithm is fast and takes (at most) hours to build about 20 models per proteins. Assessment of final models is done by a comprehensive score which is a combination of signals designed by a similar technique to the approach we developed for template detection.

Prediction of Functional Sites in Predicted Protein Structures Using Dynamics Perturbation Analysis

J. D. Cohn[1], M. E. Wall[1,2,3]
[1] Computer, Computational, and Statistical Sciences Division, [2] Bioscience Division, and [3] Center for Nonlinear Studies
Los Alamos National Laboratory, Los Alamos, NM 87505 USA
mewall@lanl.gov

Dynamics perturbation analysis (DPA)[1-3] finds regions in a protein structure where proteins are "ticklish," *i.e.*, where interactions cause a large change in protein dynamics. Such regions were shown to predict the locations of native binding sites in a docking test set[3]. Recently, we showed that an accelerated algorithm, Fast DPA, also predicts the locations of native binding sites[4]. Fast DPA is highly scalable and takes only a few minutes to predict functional sites in a typical protein domain. We have also demonstrated its use for high-throughput prediction of functional sites in 50,000 protein domains[5]. Although we have demonstrated that Fast DPA can be used as a part of genome-wide protein structure prediction pipelines, the performance of the method for prediction of functional sites in predicted structures was untested.

We submitted 127 Fast DPA predictions for evaluation in the CASP8 human function prediction experiment. Our implementation of Fast DPA was similar to a previous application[4]. Given an input PDB structure, MSMS[6] was run to generate test points on the surface of the protein. Protein vibrations were modeled using an Elastic Network Model[7], with a cutoff distance for interactions between protein Cα atoms of 8.5 Å. The cutoff distance between a test point and the protein was 14 Å, and the interaction strength between a test point and protein atoms was 12 times the strength of the interaction between protein atoms. For each test point, the relative entropy $D_{\mathbf{x}}$ was calculated between protein conformational distributions with and without the test point interaction.

To predict functional sites, the distribution of $D_{\mathbf{x}}$ values was fit to an extreme value distribution. Points with $D_{\mathbf{x}}$ values in the upper 4% of the distribution were selected and spatially clustered using the OPTICS algorithm[8] with a distance threshold of 6 Å and a minimum of 3 points per cluster. C_α atoms within 6 Å of any point in a cluster were selected and were used to define predicted functional sites.

Although our predictions were submitted as human function predictions, we used an automated method, which is more in the spirit of the server prediction category. We implemented Fast DPA as in Ref. [4] to analyze models 1-5 predicted by the BAKER-ROBETTA server for the CASP8 experiment. Predictions in CASP format were prepared by listing residues from the predicted functional sites for model 1 in the "Binding site" field. In addition, we counted the total number models for which each residue was included in the predictions. The resulting counts were listed in the "Comment" field as a measure of confidence for each predicted residue in a binding site.

1. Ming D. & Wall M.E. (2005). Quantifying allosteric effects in proteins. Proteins 59,697-707.
2. Ming D. & Wall M.E. (2005). Allostery in a coarse-grained model of protein dynamics. Phys Rev Lett 95,198301.
3. Ming D. & Wall M.E. (2006). Interactions in native binding sites cause a large change in protein dynamics. J. Mol. Biol. 358,213-223.
4. Ming D., Cohn J.D. & Wall M.E. (2008). Fast dynamics perturbation analysis for prediction of functional sites. BMC Structural Biology 8,5.
5. Cohn J.D., Ming D. & Wall ME. (2008). Prediction of Functional Sites in SCOP Domains using Dynamics Perturbation Analysis. AFP-Biosapiens 2008, July 18--19, Toronto, Canada. Extended abstract available at http://precedings.nature.com/documents/2209/version/1
6. Sanner M.F., Olson A.J. & Spehner J.C. (1996). Reduced surface: an efficient way to compute molecular surfaces. Biopolymers 38,305-320.
7. Atilgan A.R., Durell S.R., Jernigan R.L., Demirel M.C., Keskin O. & Bahar I. (2001). Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophys J 80,505-515.
8. Ankerst M., Breunig M.M., Kriegel H.P. & Sander J. (1999). OPTICS: ordering points to identify the clustering structure. Proceedings of the ACM SIGMON International Conference on Management of Data 1999, 28,49-60.

Mariner1

Homology modeled using multiple sequence and structure alignments

H. Rangwala[1] and G. Karypis[2]
[1] - *Computer Science/Bioinformatics, George Mason University, Fairfax,VA 22030*
[2] - *Computer Science, University of Minnesota, Minneapolis, MN 55455*
rangwala@cs.gmu.edu

The basic protocol for our CASP8 server (MARINER - MinnesotA pRotein modelINg servER) was to select a set of templates using profile and secondary structure methods, perform multiple sequence as well as structure alignments, followed with a model building step.

Given a target protein sequence, we primarily use DOMPro3 to identify the possible domain boundaries, which are further verified and changed based on domain prediction results using results of several other domain prediction methods. Each predicted domain for the target, is treated individually, where we predict the secondary structure using YASSPP1. (YASSPP uses a two level kernel based method to determine the secondary structure elements for the protein sequence.).

Profile-profile alignment that incorporates both YASSPP and PSI-BLAST profiles is used to identify the top template candidates. The template database for this search was restricted to a set of 9000 non-redundant proteins (with pairwise sequence identity less than 30%). Using the identified top templates, a multiple structure alignment was constructed from the template structures (using MUSTANG). The resulting multiple structure alignment was used to induce a multiple sequence alignment. The target sequence was then aligned to the induced sequence alignment, and pairwise alignments between the set of template-target sequences were used to force the final modeling by MODELER. The target models were then refined using side-chain refinement algorithm SCWRL. The model quality was evaluated using MODELER's output. It is well known that using multiple templates lead to improvement of structure prediction results. With this server, we wanted to test a simple way of selecting multiple templates and integrating with a multiple structure alignment program. The potential limitation of this server is to have a single protocol for all target proteins, irrespective of the hardness of the target. In the future, we intend to refine our selection scheme and improve the sensitivity of the multiple structure alignment program by additional constraints, specific to modeling.
We also tried selection of templates using direct profile-based kernel methods2,3. We classify each of the domain sequences in one of the 945 fold classes obtained from the SCOP (Version 1.69) database. We specifically use the Smith-Waterman profile based kernel function with optimized parameters to learn 945 one-versus-rest discriminatory models. For every target domain sequence we classify the domain into one of the fold classes. To allow for possibility of errors at this classification stage we pick the three top scoring fold models. Having selected the top scoring fold classes, which by definition have a tendency to share similar protein structure irrespective of the sequence identity, we select a set of templates from these fold classes. These templates are selected with Smith-Waterman alignment algorithm using position specific scoring matrices.

1. Karypis, G. (2006). YASSPP: Better Kernels and Coding Schemes Lead to Improvements in SVM-based Secondary Structure Prediction. Proteins: Structure, Function and Bioinformatics. 64(3), 575-586.
2. Rangwala, H.S. and Karypis, G. (2005). Profile Based Direct Kernels for Remote Homology Detection and Fold Recognition. Bioinformatics. 31(23), 4239-4247.
3. Cheng, J., Sweredoski, M., and Baldi, P. (2006). DOMpro: Protein Domain Prediction Using Profiles, Secondary Structure, Relevant Solvent Accessibility, and Recursive Neural Networks. Data Mining and Knowledge Discovery. 13(1), 1-10.
4. Rangwala, H. S. and Karypis, G. (2006). Building Multiclass Classifiers for Remote Homology Detection and Fold Recognition. BMC Bioinformatics (under review).

Disordered Region and Functional Site Prediction Using MONSTER/PROSAT

H. Rangwala[1] and G. Karypis[2]
[1] - *Computer Science/Bioinformatics, George Mason University, Fairfax,VA 22030*
[2] - *Computer Science, University of Minnesota, Minneapolis, MN 55455*
rangwala@cs.gmu.edu

Availability: http://bio.dtc.umn.edu/monster
http://bio.dtc.umn.edu/prosat

We predicted residues of the protein to be in the disordered region prediction or being functionally active (i.e., have a tendency to bind to small molecules or ligands) using PSI-BLAST derived profile information for a local set of residues within a discriminatory learning framework.

For effective development and training of residue-wise prediction models, we have developed a general purpose protein residue annotation toolkit called PROSAT. This toolkit uses a support vector machine framework and is capable of predicting both a discrete label or a continuous value. PROSAT allows use of any type
of sequence information with residues for annotation. For every residue, PROSAT encodes the input information from the residue and its neighbors. We introduce
a new flexible encoding scheme that differentially weighs information extracted from neighboring residues, based on the distance to the central residue. PROSAT also uses an exponential second-order kernel function shown to be effective in capturing pair-wise interactions between residues, and hence improve the classification and regression performance for the annotation problems.

To the best of our knowledge, PROSAT is the first tool that is designed to allow life science researchers to quickly and efficiently train SVM-based models for annotating protein residues with any desired property. The kernel functions implemented are also optimized for speed, by utilizing fast vector-based operation routines within the CBLAS library.

The dataset used for training the disordered region prediction model was identical to the one used for the DisPro[3] program. This dataset consisted of 723 sequences (215612 residues) with the maximum pairwise sequence identity being 30%. The ligand-binding prediction model was trained using 400 sequences derived from the PDBBind[4] database with maximum pairwise sequence identity being 40%. This dataset was used in our recent work on homology modeling of ligand-binding specific regions[1]. Predictions for the disordered prediction, and ligand binding predictions can be accessed via our web server called MONSTER (http://bio.dtc.umn.edu) or by training models using the PROSAT toolkit.

MONSTER is a server for predicting the local structure and function properties of protein residues. MONSTER provides residue-wise annotation services that include
secondary structure, transmembrane-helix region, disorder region, protein-DNA binding site, ligand-binding site, local structure alphabet, solvent accessibility
surface area, and residue-wise contact order prediction.

1. Kauffman C., Rangwala, H., and Karypis, G. (2008). Improving Homology Models for Protein-Ligand Binding Sites. Proceedings LSS Computational Systems Biology Conference. 211-222.
2. Rangwala H., Kauffman C., and Karypis G. (2007). A generalized framework for protein sequence annotation. Proceedings of the NIPS workshop Machine Learning in Computational Biology, Whistler, Canada. Available at http://www.cs.umn.edu/~karypis
3. J. Cheng, M. Sweredoski, and P. Baldi. Accurate Prediction of Protein Disordered Regions by Mining Protein Structure Data, Data Mining and Knowledge Discovery, vol. 11, no. 3, pp. 213-222, 2005
4. Wang, R.; Fang, X.; Lu, Y.; Yang, C.-Y.; Wang, S. "The PDBbind Database: Methodologies and updates", J. Med. Chem., 2005; 48(12); 4111-4119.

McGuffin

Manual predictions in the disorder prediction, quality assessment and tertiary structure prediction categories

L.J. McGuffin
School of Biological Sciences, University of Reading, Whiteknights, Reading RG6 6AS, UK
l.j.mcguffin@reading.ac.uk

Manual predictions were submitted in the disorder prediction (DR) category using the DISOclust method[1]. The same protocol used for the DISOclust server predictions was followed, however, in this case all server models were used to produce predictions rather than just those obtained from the nFOLD3 server. For each target, the tarball containing all CASP8 server models was submitted to the ModFOLD server[2] and the DISOclust option was selected. The resulting predictions were then uploaded using the CASP8 manual submission form.

Quality assessment (QMODE1) was carried out for all server models using ModFOLD version 2.0. This novel method attempts to combine the best features of the ModFOLD and ModFOLDclust methods[2,3]. ModFOLDclust is more accurate for ranking multiple models than ModFOLD, however it cannot produce a score for a single model as it relies on clustering with other models. Conversely, whilst ModFOLD can produce a score for a single model it is generally less accurate and does not provide per-residue accuracy predictions. Version 2.0 of the ModFOLD method combines the 6 scores from the ModFOLD method with the clustering score from ModFOLDclust using an artificial neural network, producing a single score for each model. The beta version of the ModFOLD 2.0 server is now being developed, which is able to carry out QMODE2 predictions on either single or multiple models. Each model is firstly analysed using ModFOLD version 1.1 and then it is compared with models obtained from the nFOLD3 server using the ModFOLDclust method. The combined prediction scores are then returned to the user.

For the tertiary structure (TS) category, manual predictions were made purely using ModFOLD version 2.0 for model selection. The top five server models, according the ModFOLD 2.0 ranking, were selected and submitted as TS predictions. The only modifications made to the models were in cases where the full backbone did not exist, in which case the program BBQ[4] was used to reconstruct the chain. In addition, for each model the ModFOLDclust predicted per-residue error was added into the B-factor column for each set of ATOM records.

1. McGuffin,L.J. (2008) Intrinsic disorder prediction from the analysis of multiple protein fold recognition models. Bioinformatics. 24, 1798-1804.
2. McGuffin,L.J. (2008) The ModFOLD Server for the Quality Assessment of Protein Structural Models. Bioinformatics. 24, 586-587.
3. McGuffin,L.J. (2007) Benchmarking consensus model quality assessment for protein fold recognition, BMC Bioinformatics. 8, 345.
4. Gront,D., Kmiecik,S., Kolinski,A. (2007) Backbone building from quadrilaterals: a fast and accurate algorithm for protein backbone reconstruction from alpha carbon coordinates. J Comput. Chem. 28, 1593-1597.

MeilerLabRene

De novo tertiary structure prediction from secondary structure elements using Monte Carlo and knowledge based potentials

N. Woetzel[1], R.D. Starizbichler[1], N.S. Alexander[1], M. Karakas[1], J. Koehler[1], S. Lindert[1], R. Mueller[1], M. Butkiewicz[1] and J. Meiler[1*]
[1] – *Vanderbilt University, Department of Chemistry, Nashville, TN, 465 21st Ave South BIOSCI MRBIII*
jens@jens-meiler.de

Novel fold topologies that are not yet represented in the PDB are often found in large macromolecular complexes or membrane proteins outside individual soluble domains. Structure determination is challenging for such proteins as many of these systems evade crystallization, alternative approaches such as cryo-electron microscopy or EPR spectroscopy yield low-resolution or sparse data sets, and these systems are too large for computational *de novo* protein structure prediction. The presented algorithm determines tertiary structures from sequence by assembling predicted secondary structure elements (SSEs) of α-helices and β-strands in space. This method seeks to overcome size and complexity limits of previous approaches

by discontinuing the amino acid chain in the folding simulation and limiting the sampling of flexible loop regions. Employing a Monte Carlo[1] procedure, the sampling trajectory is guided by knowledge based potentials that evaluate amino acids' pair interaction and environment, SSE packing, loop closure, and protein compactness. The method is tailored to be used in conjunction with low-resolution or sparse experimental data sets which often provide more readily restraints for regions of defined secondary structure.

The method was used to produce models for given target sequences. The submission candidates were pre selected from these models based on low energy and formation of tight clusters. This was followed by model completion protocol using Rosetta[2] to add loops and side chains to the submission candidates.

Three secondary structure prediction methods, PSIPRED[3], SAM[4] and JUFO[5], have been equally weighted to achieve a consensus three state secondary structure prediction. Stretches of sequence with consecutive α-helix or β-strand predictions above a given threshold were identified as α-helical and β-strand SSEs.

The predicted SSEs were then passed to the assembly protocol. The assembly method is a simulated annealing Monte Carlo minimization employing the Metropolis criterion[6]. The following Monte Carol steps are utilized to generate new protein models throughout the minimization process: *SSEAdd:* A selected SSE from the predicted SSEs is added to the protein model. *SSERemove:* A selected SSE is removed from the protein model. *SSESwap:* Two selected SSEs are exchanged. *SSERotate:* A selected SSE is rotated along an internal axis. *SSETranslate:* A selected SSE is translated along an internal axis. *SSETransform:* A small rotation and translation is applied to a SSE in a single step. *SSEFlip:* A selected SSE is inverted in the direction of its main axis. One trajectory consists of 10,000 steps.

The assembly protocol was used to create 50,000 models. From this a group of submission candidates were selected. The best model by energy and the model closest to the top ranked BIOINFO[7] homology model by C-alpha RMSD were immediately chosen. In addition the best 10,000 models by energy were clustered using the statistical package R[8]. Clusters were calculated according to C-alpha RMSD100[9] using average linkage. Between ten and twenty of the largest clusters were chosen, and the cluster center and the lowest energy model from each cluster were selected for the loop building and high resolution refinement. A last candidate was determined by running mammoth[10] over models from the selected clusters against a ~1800 structure database (culled by PISCES[11]) and choosing the best by Z-score.

The models have been completed by applying Rosetta's loop building protocol which generated up to 1,000 models for each pre selected model. All models were subjected to high resolution refinement by undergoing eight iterations of alternating gradient based backbone minimization and side chain repacking in Rosetta. From the resultant models for each candidate, the best model by Rosetta score was selected for final submission. A fifth model was chosen at the submitter's discretion.

1. Metropolis, N. & Ulam, S. (1949). The Monte Carlo method. J Am Stat Assoc 44, 335-41.
2. Bradley, P., Malmstrom, L., Qian, B., Schonbrun, J., Chivian, D., Kim, D. E., Meiler, J., Misura, K. M. & Baker, D. (2005). Free modeling with Rosetta in CASP6. Proteins 61 Suppl 7, 128-34.
3. Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. Journal of Molecular Biology 293, 195-202.
4. Karplus, K., Katzman, S., Shackleford, G., Koeva, M., Draper, J., Barnes, B., Soriano, M. & Hughey, R. (2005). SAM-T04: What is new in protein-structure prediction for CASP6. Proteins-Structure Function and Bioinformatics 61, 135-142.
5. Meiler, J. & Baker, D. (2003). Coupled prediction of protein secondary and tertiary structure. Proc Natl Acad Sci U S A 100, 12105-10.
6. Metropolis, N. R., A.; Rosenbluth, M.; Teller A. . (1953). Equations of state calculations by fast computing machines. Journal of Chemical Physics 21, 1087 - 1091.
7. Ginalski, K., Elofsson, A., Fischer, D. & Rychlewski, L. (2003). 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 19, 1015-8.
8. Team, R. D. C. (2008). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
9. Carugo, O. & Pongor, S. (2001). A normalized root-mean-square distance for comparing protein three-dimensional structures. Protein Science 10, 1470-1473.
10. Ortiz, A. R., Strauss, C. E. M. & Olmea, O. (2002). MAMMOTH (Matching molecular models obtained from theory): An automated method for model comparison. Protein Science 11, 2606-2621.
11. Wang, G. L. & Dunbrack, R. L. (2005). PISCES: recent improvements to a PDB sequence culling server. Nucleic Acids Research 33, W94-W98.

Metaprdos

Prediction of disordered regions in proteins based on meta approach

T. Ishida[1] and K. Kinoshita[1]
[1] – *Human Genome Center, Institute of Medical Science, the University of Tokyo*
t-ishida@hgc.jp

We predicted disordered regions in proteins by using meta prediction system named metaPrDOS[1]. This prediction system comprises two main steps. In the first step, an input sequence is submitted to each disorder predictor, and prediction results from all predictors are collected. We used seven predictors: PrDOS[2], DISOPRED2[3], DisEMBL[4], DISPROT (VSL2P)[5], DISpro[6], IUpred[7] and POODLE-S[8]. Each predictor will perform its own prediction for each residue, and the result is obtained as a disorder tendency. In the second step, the meta predictor integrates the prediction results and determines the disorder tendency for each residue. Thus, the dimension of the input vector for meta predictor corresponds to the number of component predictors. We adopted the support vector machine (SVM) as the meta predictor. Finally, the decision value of the SVM is scaled from 0.0 to 1.0, and it is returned as a prediction result.

1. Ishida, T. and Kinoshita, K. (2008) Prediction of disordered regions in proteins based on the meta approach, *Bioinformatics* **24**, 1344-1348.
2. Ishida, T. and Kinoshita, K. (2007) PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Res.*, **35**, W460–W464
3. Jones, DT and Ward, JJ. (2003) Prediction of disordered regions in proteins from position specific score matrices. *Proteins*, **53**, 573–578.
4. Linding R, et al. (2003) Protein disorder prediction: implications for structural proteomics. *Structure*, **11**, 1453–1459
5. Peng K, et al. (2006) Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics*, **7**, 208.
6. Cheng JL, et al. (2005) Accurate prediction of protein disordered regions by mining protein structure data. *Data Min. Knowl. Discov.*, **11**, 213–222.
7. Shimizu K, et al. (2007) POODLE-S: web application for predicting protein disorder by using physicochemical features and reduced amino acid set of a position-specific scoring matrix. *Bioinformatics*, **23**, 2337–2338.

METATASSER

METATASSER: A 3D-jury threading approach with TASSER model assembly/refinement

S. B. Pandit, H. Zhou and J. Skolnick
Center for the Study of Systems Biology, Georgia Institute of Technology,
250 14th Street, N.W., Atlanta, GA 30318
skolnick@gatech.edu

METATASSER employs the 3D-jury[1] approach to select threading templates from SPARKS2[2], SP3[3] and PROSPECTOR_3[4], which provide aligned fragments and tertiary restraints as an input to TASSER[5]. In our implementation of the 3D-jury approach, the ten top-scoring templates from each threading methods are compared with each other using the structural alignment method TM-align[6] with the TM-score[7] used as the similarity measure. For Medium/Hard targets, in addition to the tertiary restraints from templates, restraints derived from predicted supersecondary structures chunks are used[8]. In TASSER[5], the template derived continuous fragments blocks are kept rigid and are off-lattice to retain their geometric accuracy, while unaligned regions are modeled on a cubic lattice by an *ab initio* procedure and serve as linkage points for rigid body fragment rotations. Parallel Hyperbolic Monte Carlo (MC) sampling (PHS)[9] is used to explore conformational space by rearranging the continuous fragments excised from the templates. Conformations are selected using an optimized force field that includes knowledge-based statistical potentials describing short-range backbone correlations, pairwise interactions, hydrogen-bonding, secondary structure propensities, and consensus contact restraints. Multiple independent TASSER simulations are performed for each target sequence and structures are clustered using SPICKER[10]. The top five cluster centroids selected from each of the simulations are ranked using TASSER-QA[11]. The top five ranked cluster centroids are submitted as final models after building the side-chains using PULCHRA[12].

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. 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.
3. Zhou, H. & Zhou, Y. (2005) Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins 58, 321-328.
4. Skolnick, J., Kihara, D., & Zhang. Y. (2004) Development and large scale benchmark testing of the PROSPECTOR_3 threading algorithm. Proteins 56, 502-518.
5. Zhang, Y. & Skolnick, J. (2004) Automated structure prediction of weakly homologous proteins on a genomic scale. Proc. Natl. Acad. Sci. USA. 101, 7594-7599.
6. Zhang, Y., & Skolnick, J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 33, 2302-2309.
7. Zhang, Y., & Skolnick, J. (2004) Scoring function for automated assessment of protein structure template quality. Proteins. 57, 702-710.
8. Zhou, H., & Skolnick, J. (2007) Ab initio protein structure prediction using chunk-TASSER. Biophysical Journal 93, 1510-1518.
9. Zhang, Y., Kihara, D., & Skolnick, J. (2002) Local energy landscape flattening: parallel hyperbolic Monte Carlo sampling of protein folding. Proteins. 48, 192-201.
10. Zhang, Y., & Skolnick, J. (2004) SPICKER: a clustering approach to identify near-native protein folds. J. Comput. Chem. 25, 865-871.
11. Zhou, H., & Skolnick, J. (2007) Protein model quality assessment prediction by combining fragment comparisons and a consensus Cα contact potential. Proteins. 71, 1211-1218.
12. Rotkiewicz, P., & Skolnick, J. (2008) Fast procedure for reconstruction of full-atom protein models from reduced representations. Journal of Computational Chemistry 29, 1460-1465.

MicrotechNano

Protein Folding Shape Alignment – New Method for Quality Assessment of Protein Structure Prediction

J. Yang
MicrotechNano, LLC, Indianapolis, Indiana 46234, USA
jiaan@proteinshape.com

Recently developed Protein Folding Shape Code (PFSC) is a powerful tool for protein structure comparison.[1] In the PFSC approach, a set of 27 PFSC vectors is mathematically derived from an enclosed space mapping all possible folding shapes of five successive C□ atoms. With PFSC, any protein 3-D structure is able to be described with one-dimensional string, which represents the changes of folding structures from N-terminus to C-terminus along protein backbone, including regular secondary and irregular tertian structures.

Based on the PFSC, the protein structures are able to be compared with alignment of one-dimension folding shape strings, which is called as Protein Folding Shape Alignment (PFSA). In the PFSA, all global and local structures of protein will treated with equal weight. The algorithm of alignment will be utilized to evaluate the similarity or dissimilarity of protein structures. Therefore, ambiguous procedure of superposition of 3D structures will be avoided. The similarity score will be obtained as global measurement and the detail local structure comparison will be displayed by one-dimension folding shape strings for each residue.

With PFSA, the predicted protein structures are compared with the target protein in CASP8. The score of PFSA ranks predicted protein structures according with structural global similarity. The local structures are able to be compared with PFSC alignment table, and the structural similarity and dissimilarity are able to be displayed in detail on residue-residue level.

1. Yang, J. Comprehensive description of protein structures using protein folding shape code. Proteins 2008;71.3:1497-1518.

Structure prediction combining the template-based RAPTOR algorithm with the ItFix *ab initio* method.

J. DeBartolo[1], G. Hockey[2], F. Zhou[6], J. Peng[6], A. Augustyn[2], A. Adhikar[2], J. Xu[6], K. F. Freed[2,3,4] and T. R. Sosnick[1,2,5]

[1]Dept. of Biochemistry and Molecular Biology, [2]Dept. of Chemistry, [3]The James Franck Inst., [4]Computation Inst., [5]Inst. for Biophysical Dynamics, The University of Chicago, [6]Toyota Technology Inst. at Chicago.

trsosnic@uchicago.edu

Our goal during the CASP8 experiment was to combine "ItFix" modeling tools to refine and corroborate template-based models created using RAPTOR1 and when no templates were available, to generate ab initio structures. The modeling tools include a Ca-level statistical potential, trimer sampling, and iterative fixing for the prediction of secondary structure. In the ItFix algorithm, secondary and tertiary prediction is integral to and a consequence of the folding process. Hence, the algorithm may share some benefits that real proteins gain by folding along a robust and efficient pathway. When RAPTOR produced multiple models of similar energy, the ItFix ab initio folding method provided additional information for selecting the best model by comparing the secondary structure and 3D contacts determined from ab initio simulations with the respective secondary structure and contacts of each RAPTOR-generated model. An important aspect of our strategy was to analyze the RAPTOR structures and determine if and how much ab initio folding was necessary for each target sequence, thereby maximizing the utilization efficiency of available resources. We could then determine whether a template-based model needed some specific refinement or whether ab initio folding was necessary for structure prediction.

The ItFix refinement protocol for refining the RAPTOR model focused primarily on regions where the template modeling was presumed to be least accurate (e.g., at insertions). The refinement operation typically consisted of breaking the chain inside the misaligned region and then subjecting that same region to ab initio folding with a constraining term added to the energy function to reconnect the broken ends of the chain. Importantly, the ItFix conformation-sampling tools allowed a high level of flexibility in sampling chain conformations. For example, if the confidence in the secondary structure prediction was high, the backbone sampling was made contingent on that secondary structure for maximal efficiency of conformational search, while allowing flexibility when warranted.

Allowing flexibility in the ItFix conformational sampling protocol was equally useful when RAPTOR was unable to produce a template-based model for a given sequence, and the ItFix ab initio folding tools were necessary. When the PSIPRED secondary structure prediction for the target sequence yielded a very high confidence prediction at a given position, we fixed the conformational sampling at that position to retain the predicted secondary structure, while leaving all other positions open to search all secondary structure types. The ItFix ab initio folding algorithm then allowed the unknown secondary structures to be determined through successive rounds of folding simulations. This iterative process produced a folding-enhanced secondary structure for the given sequence, along with an ensemble of tertiary structures from which a predicted tertiary structure was chosen.

Our statistical potential includes only main chain heavy atoms and side chain Ca atoms2. In addition to amino acid type, the statistical potential depends on secondary structure and side chain orientation. This energy function is minimized using a Monte Carlo Simulated Annealing algorithm in which the elementary moves involve the sampling of one pair of □□□□backbone torsional angles at randomly chosen positions. These angles are selected from a library of trimers that match the target amino acid sequence and that satisfy any specified secondary structure. To enhance the number of trimers, we generate a multiple sequence alignment (MSA) for the target sequence by performing a BLAST search of the protein sequence database for amino acid sequences that are homologous to the target. Amino acids types that occur frequently at a position in the MSA are added to a substitution matrix of allowed amino acids at that position. This process increases the diversity of angles to be sampled and also improves the distribution of secondary structures within the fragment library. As such, the use of MSA's is an invaluable enhancement to the accuracy of both secondary and tertiary structure predictions.

1. Xu, J., Li, M., Kim, D., and Xu, Y. (2003) RAPTOR: optimal protein threading by linear programming. J. Bioinform. Comp. Biol. 1, 95-117.
2. Fitzgerald, J. E., Jha, A. K., Colubri, A., Sosnick, T. R., and Freed, K. F. (2007) Reduced Cbeta statistical potentials can outperform all-atom potentials in decoy identification. Protein Sci. 16, 2123-2139.

ModFOLD
ModFOLDclust

Model quality assessment using the ModFOLD server

L.J. McGuffin
School of Biological Sciences, University of Reading, Whiteknights, Reading RG6 6AS, UK
l.j.mcguffin@reading.ac.uk

Predictions were submitted in the quality assessment (QA) category using the ModFOLD server[1]. The server includes two different methods: ModFOLD, which was used for QMODE1 predictions and ModFOLDclust which was used for QMODE2 predictions.

The original ModFOLD protocol combined scores obtained from the ModSSEA method[2], the MODCHECK method[3] and the two ProQ methods[4] using a neural network trained with the TM-score[5]. The latest implementation of the ModFOLD method (version 1.1) has been re-trained and now includes two additional secondary structure scores, similar to those used by Eramian and colleagues[6], as inputs to the neural network.

The server also includes an option for clustering multiple models using the ModFOLDclust method. The method carries out pairwise comparisons of models in order to produce both global and local predictions of model accuracy. The global clustering score is based on the 3D-Jury method[7], whereby each model is compared to every other model and the average structural similarity score is calculated. However, in this application, the TM-score is used for pairwise comparisons, with a score cut-off of >0.2. This emulation of the 3D-Jury score has been previously benchmarked on the set of CASP7 server models and was shown to significantly outperform every method tested for the selection of the highest quality models[2].

In addition to the global clustering score, the ModFOLDclust method incorporates the scoring of the local model quality on a per-residue basis. The local model quality is evaluated by using a score similar to the average *S*-score[8], which was originally used for model evaluation in the 3D-SHOTGUN method [9] and was more recently benchmarked using the Pcons server[10]. The idea in this implementation is to reuse each pairwise model superposition, carried out in the calculation of the global score, in order to evaluate the local structural conservation of each residue. Here, the *S*-score is used to evaluate residues that are within 3.9Å according to pairwise TM-score superpositions, where the TM-scores >0.2. The S-score is defined as: $S_i = 1/(1+(d_i/d_0)^2)$, where S_i ranges from 0 to 1, d_i is the distance between structurally aligned residues and d_0 is the distance threshold (3.9). An S_i score of 0 is given if d_i >3.9Å. The *S*-scores for each residue are then summed and the mean score is taken. The mean *S*-score for each residue is then converted to the predicted distance from the native structure, by simply rearranging the equation: $d_i = d_0\sqrt{((1/S_i)-1)}$.

The ModFOLD web server is available at the following URL: http://www.reading.ac.uk/bioinf/ModFOLD/

1. McGuffin,L.J. (2008) The ModFOLD Server for the Quality Assessment of Protein Structural Models. Bioinformatics. 24, 586-587.
2. McGuffin,L.J (2007) Benchmarking consensus model quality assessment for protein fold recognition, BMC Bioinformatics. 8, 345.
3. Pettitt,C.S., McGuffin,L.J. & Jones,D.T. (2005) Improving sequenced based fold recognition by use of 3D model quality assessment. Bioinformatics. 21, 3509-3515.
4. Wallner,B. & Elofsson,A. (2003) Can correct protein models be identified? Protein Sci. 12, 1073-1086.
5. Zhang,Y. & Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality. Proteins. 57, 702-710
6. Eramian,D., Shen,M.Y., Devos,D., Melo,F., Sali,A. & Marti-Renom,M.A. (2006) A composite score for predicting errors in protein structure models. Protein Sci. 15, 1653-1666.
7. Ginalski,K., Elofsson,A., Fischer,D. & Rychlewski,L. (2003) 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics. 19, 1015-1018.
8. Levitt,M. & Gerstein, M. (1998) A unified statistical framework for sequence comparison and structure comparison, Proc Natl Acad Sci U S A 95, 5913-5920.
9. Fischer, D. (2003) 3D-SHOTGUN: a novel, cooperative, fold-recognition meta-predictor. Proteins. 51, 434-441.

10. Wallner,B. & Elofsson,A. (2006) Identification of correct regions in protein models using structural, alignment, and consensus information, Protein Sci. 15, 900-913.

MUFOLD-MD

Selection of Near-native Structures by Means of Molecular Dynamics simulations

B. Barz[1], Q.Wang[2], J. Zhang[2,3], Z. He[2,3], D. Xu[2,3], Y. Shang[2], I. Kosztin[1]
[1]Department of Physics and Astronomy, [2]Department of Computer Science [3]Christopher S. Bond Life Sciences Center, University of Missouri,
Columbia, MO 65211, USA
KosztinI@missouri.edu

The correct identification of near-native structures from a large pool of previously generated decoys is an important step in most protein structure prediction methods. In the case of globular proteins one expects that the closer the predicted structure to the native one (i.e., the smaller the corresponding RMSD) the higher its stability. Thus, the quantitative assessment of the relative stability of the predicted protein structures, e.g., against gradual heating by all-atom molecular dynamics (MD) simulations, provides an alternative for ranking the quality of these structures. We have used this approach to develop and implement the MD-Ranking (MDR) method. MDR was tested by us in the CASP8 competition as part of our MUFOLD-MD server.

Our MUFOLD-MD server for predicting 3D structures of a target protein, using as input its primary structure (amino acid sequence), has two interconnected modules. The 1st one uses the publicly available Rosetta software[1-4] to generate up-to ten thousands structures (using the *ab-inition* method) from which the 64 lowest energy structures are retained. The 2nd module contains the implementation of our MDR method, which is used to select the top 5 structures from the models returned by the 1st module.

The decoy generation in the 1st module is preceded by gathering secondary structure information from the amino acids sequence using the PSIPRED method[5] and by building fragment libraries from the NCBI database files. The main decoy generation code was compiled on a 64-bit operating system with the open-source GCC compiler. We have used 32 dual-core Intel Xeon EM64T-2.8GHz CPUs to generate 10000 low resolution (backbone atoms) structures. We found that due to the parallel nature of the decoy generation procedure, many of the structures built on different CPUs were identical; in some cases the total number of unique structures was reduced from 10000 to less than 2000. The generated decoys are further filtered with the Rosetta energy function by selecting the top 64 distinct structures, with lowest energy. These structures are used as input for the MDR method.

Our MDR method consists of several important steps. First, an all-atom, high-resolution structure is built for each of the 64 predicted low resolution structures. For this, the coordinates of the missing side-chain heavy atoms are determined by using the program PULCHRA[6], and the hydrogen atoms are added by using PSFGEN, which is part of the visual molecular dynamics (VMD) package[7]. Next, the obtained structures are optimized by removing the bad contacts through energy minimization. Finally, the stability of the structures is tested by monitoring the change of their RMSD (with respect to their low-resolution structures) during the MD simulation of their scheduled heating at a rate of 1 K/ps. The MD simulations are carried out in vacuum by coupling the system to a Langevin heat bath whose temperature can be varied according to a desired protocol. All energy minimization and MD simulations were performed by employing the CHARMM force field[8, 9] and the parallel NAMD2.6 MD simulation program[10]. Based on extensive testing of the MDR method we have found that statistically the best ranking parameter of the predicted structures is their mean RMSD during heating from 40K to 140K. This can be achieved through 100ps long MD simulations that take a matter of hours on a single dual core Intel Xeon EM64T-2.8GHz CPU.

It should be noted that the success of the MDR method, for ranking the predicted low-resolution structures, relies heavily on how well the high-resolution (all-atoms) structures are constructed by PULCHRA and PSFGEN. Also, the omission of the explicit solvent in the MD simulations may influence the final ranking of the structures.

The MUFOLD-MD server was used for protein structure prediction in the CASP8 competition. Once the native structures for the CASP8 targets were released we were able to assess the quality of our predicted structures and the efficiency of the MDR method. The results of this analysis will be presented during the CASP8 meeting.

1. Bonneau, R., Strauss, C. E. M., Rohl, C. A., Chivian, D., Bradley, P., Malmström, L., Robertson, T. & Baker, D. (2002) Journal of Molecular Biology 322, 65-78.
2. Bonneau, R., Tsai, J., Ruczinski, I., Chivian, D., Rohl, C., Strauss, C. E. M. & Baker, D. (2001) Proteins: Structure, Function, and Genetics 45, 119-126.
3. Simons, K. T., Ingo Ruczinski, Kooperberg, C., Fox, B. A., Bystroff, C. & Baker, D. (1999) Proteins: Structure, Function, and Genetics 34, 82-95.
4. Simons, K. T., Kooperberg, C., Huang, E. & Baker, D. (1997) Journal of Molecular Biology 268, 209-225.
5. Jones, D. T. (1999) Journal of Molecular Biology 292, 195-202.
6. Rotkiewicz, P. & Skolnick, J. (2008) Journal of Computational Chemistry 29, 1460-1465.
7. Humphrey, W., Dalke, A. & Schulten, K. (1996) J. Mol. Graphics 14, 33-38.
8. MacKerell Jr, A. D., Bashford, D., Bellott, M. & others (1992) FASEB J. 6, A143-A143.
9. MacKerell Jr, A. D., Bashford, D., Bellott, M. & others (1998) J. Phys. Chem. B 102, 3586--3616.
10. Phillips, J. C., Braun, R., Wei Wang, Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kalé, L. & Schulten, K. (2005) Journal of Computational Chemistry 26, 1781-1802.

MUFOLD-QA

Selection of Near-native Structures by Machine Learning Methods

Q. Wang[1], J. Zhang[1,3], Z. He[1,3], B. Barz[2], I. Kosztin[2], D. Xu[1,3] and Y. Shang[1]
[1]Department of Computer Science, [2]Department of Physics and Astronomy, [3]Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211, USA
shangy@missouri.edu

In protein tertiary structure prediction, it is a crucial step to select near-native structures from a large number of candidate structural models. Despite much effort to tackle the problem of protein structure selection, the discerning power of current methods is still unsatisfactory. We have used machine learning techniques to develop a new ranking method, which has improved ability of differentiating good protein structures from bad ones. Our ranking method has been implemented as our MUFOLD-QA server and tested by us in the CASP8 QA prediction.

The main idea of MUFOLD-QA is to apply machine learning methods based on the values of various energy and scoring functions for the candidate structures. This is similar to the consensus approach, which was successful in previous CASP competitions. In the model quality assessment (QA) category in CASP7, the accuracy of structure evaluation achieved by the consensus approach was consistently better than other methods[1]. The scoring functions we used include OPUS[2], Model Evaluator, Rapdf[3], Dfire energy[4], Hopp score[5], and a geometric potential[6]. The scoring functions are normalized to z-scores. The consensus method that we implemented using machine learning methods is more sophisticated and usually better than individual energy or scoring functions.

The ranking generated by our MUFOLD-QA server for predicted structures had been optimized for the purpose of structural model selection. Although we had not targeted at the requirement of CASP8 QA prediction initially, we modified our method and submitted our predictions. We will present some assessments of our method in comparison with others at the CASP8 conference.

1. Cozzetto D., Kryshtafovych, A., Ceriani M. and Tramontano, A. Assessment of predictions in the model quality assessment category. Proteins: Structure, Function, and Bioinformatics, 69 Suppl 8:175-83, 2007.
2. Wu, Y., Lu, M., Chen, M., Li J. and Ma, J. OPUS-Ca: A knowledge-based potential function requiring only Ca positions. Protein Science, 16:1449-1463, 2007.
3. Samudrala, R. and Moult, J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J. Mol. Biol. 275:895–916, 1998.
4. Zhou, H. and Zhou, Y.. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Science, 11:2714-2726, 2002.
5. Sims, G.E., Kim, S.-H.. A method for evaluating the structural quality of protein models by using HOPP scoring. Proc. of the Nat. Acad. Sciences, 103(12):4428-4432, 2006.
6. Li, X. and Liang, J. Geometric packing potential function for model selection in protein structure and protein-protein binding predictions. Proteins: Structure, Function, and Bioinformatics, in press.

Model Ranking, Combination, Refinement and Assessment by MULTICOM Human Predictor

J. Cheng[1,2], Z. Wang[1], A. N. Tegge[2], X. Deng[1] and M. Dickinson[1]
[1] *Computer Science Department,* [2] *Informatics Institute, University of Missouri, Columbia, MO 65211, USA*
chengji@missouri.edu

MULTICOM is our human expert predictor participating in the prediction of tertiary structures, model quality assessment, domain boundaries, residue-residue contacts, and disorder regions. Unlike our MULTICOM server series that generated their own predictions, our MULTICOM human predictions started from all CASP8 server predictions. The server predictions were improved by automated ranking, combination, refinement, and some human interventions. The methods are described in each category as follows.

1. Tertiary Structure Prediction (TS)

Model Ranking: The server predictions of a target were downloaded from the CASP8 web site. The models were evaluated by our model evaluation tool ModelEvaluator[1], which assigned a predicted GDT-TS[2] score to each model. The models were ranked by their predicted GDT-TS scores.

Model Combination and Refinement. The top 50% of models were retained for combination and refinement. Each of the top five models was used as a seed model to do *global-local* model combination. The seed model was compared with all other models using the structure comparison tool TM-Score [3]. All the models that were globally similar to the seed model were chosen, i.e. > 80% of the regions in the models can be aligned with the seed model with less than 4 Å RMSD. The models and the seed model were used as templates for Modeller 7v7 [4] to generate a refined model, which was submitted to CASP. The model was essentially an averaged combination of selected models, which tends to be better than each individual model, according to our own benchmark. If no globally similar model was found for the seed model, which often happened for hard targets, the long local fragments of the models that were similar to the seed model were chosen. The minimum length of the local fragments started from 80 residues and may be reduced to make sure that some long local fragments were found. The structure of the local fragments and the seed model were combined and fed into Modeller, which generates a refined structure model. The approach may combine good regions from different models together in order to improve over the original seed models. According to our preliminary assessment, the procedure generated very good models that are sometime even better than all CASP server models. The combination method indeed can refine original input models. The combination and refinement procedure was fully automated. In rare cases, our human expert intervened the process to manually select seed models according to human insights.

2. Model Quality Assessment (QA)

The MULTICOM model quality assessment procedure has two fully automated steps. MULTICOM first downloaded all CASP8 QA server predictions, including our MULTICOM-CLUSTER, MULTICOM-CMFR, MULTICOM-REFINE, and MULTICOM-RANK. The predicted scores of the models in these predictions were averaged together to generate a consensus prediction. The consensus predicted quality score of the models were then used to rank all the models. The top five ranked models were selected as reference models for model comparison. Each model was compared against the reference models by TM-Score[3]. The GDT-TS score that resulted from the comparison with one reference model is the measure of the similarity between them. The averaged GDT-TS score over five reference models was used as the predicted *global* quality of the model.

During the comparison, a superimposition between each model and the reference model was generated. The superimposition was used to calculate the distance between the positions of a residue in the model and the reference model. The average distance over five reference models was used as a predicted *local* quality of the residue. The basic procedure of MULITCOM human quality assessment is the same as our MULTICOM-CLUSTER quality assessment, except that MULTICOM-CLUSTER used only our own model evaluation tool to rank models for model comparison.

In our preliminary assessment, we computed the global correlation between true GDT-TS scores and predicted GDT-TS scores for all the models and the per-target correlation for the models of each target. The global correlation is 0.935 and the average per-target correlation is 0.908. The average loss, the average difference of the GDT-TS score between the best models with highest real GDT-TS score and the No. 1 models ranked by the predicted GDT-TS score, was 4.6.

3. Domain Boundary Prediction (DP)

Our human domain boundary predictions were based on our server (MULTICOM-CMFR) predictions and human insights. The CASP tertiary structure models for a target were downloaded and ranked by our model quality assessment methods[1]. The top model was parsed into domains manually. The consensus prediction based on the MULTICOM-CMFR server and human parsing was submitted to CASP8.

4. Disorder Prediction (DR)

Our human disorder predictions were the consensus predictions based on CASP8 server predictions including our own MULTICOM-CMFR.

5. Residue-Residue Contact Prediction (RR)

Our human residue-residue contact predictions were the consensus predictions of CASP8 residue-residue contact prediction servers including our three servers: MULTICOM-CMFR, MULTICOM-RANK and MUProt. We tried to test if a community wide effort can improve contact predictions, which is still a very challenging problem.

1. Wang, Z., Tegge, A.N., Cheng, J. (2008). Evaluating the absolute quality of a single protein model using support vector machines and structural features. *Proteins*, in press.
2. Zemla, A. (2003). LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Research*. **31**, 3370-3374.
3. Zhang, Y., Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins*. **57**, 702-710.
4. Sali, A., Blundell, T.L. (1993). Comparative protein modeling by satisfaction of spatial restraints. *J. Mol. Bio.* **234**, 779-815.

MULTICOM-CLUSTER

Multi-Template Model Generation and Hybrid Model Quality Assessment by MULTICOM-CLUSTER

J. Cheng[1,2], Z. Wang[1], A. N. Tegge[2]

[1] *Computer Science Department,* [2] *Informatics Institute, University of Missouri, Columbia, MO 65211, USA*

chengji@missouri.edu

MULTICOM-CLUSTER is a server in our MULTICOM series. It participated in tertiary structure prediction and model quality assessment, which are described in the following two sections.

1. Tertiary Structure Prediction (TS)

MULTCOM-CLUSTER made structure predictions in four steps: (1) template-identification via *alternative* profile alignments and a machine learning method; (2) multiple-template combination; (3) model generation; (4) model ranking.

Template Identification. PSI-BLAST [1] was used to search a query protein sequence against the NCBI Non-Redundant protein sequence database to build three different kinds of sequence profiles including the Position Specific Scoring Matrix (PSSM) of PSI-BLAST, the hidden Markov model (HMM) of hhsearch [2], and the profile of COMPASS [3]. The PSSM profile, HMM, and COMPASS profile were searched against our in-house template sequence database, template HMM database, and template COMPASS profile database to identify homologous templates by PSI-BLAST, hhsearch, and COMPASS, respectively. The query-template alignments generated by PSI-BLAST, hhsearch, and COMPASS were kept in three different sets and ranked according to e-values. In addition, SPEM [4], a global profile-profile alignment tool, was used to align the query with the top 10 templates found by a sensitive machine learning fold recognition method [5]. This alignment created the fourth set of query-template alignments.

Multi-Template Combination. The most significant query-template alignment in each set was chosen and greedily combined with the rest of the alignments from the same set to form a multiple sequence alignment centered on the query sequence. This was done using a multi-template combination algorithm described in [6]. The most significant alignment was then removed and the second most significant alignment was combined with the remaining query-template alignments in order to generate a multiple sequence alignment using the same algorithm. The process was repeated up to 10 times to generate up to 10 multiple alignments in each set.

Model Generation. Each query-template alignment and the corresponding template structures were fed into Modeller 7v7 [7] to generate 10 models, among which the model with minimum Modeller energy was chosen as a predicted model. If no significant template was found by hhsearch (e-value < 10^{-3}) and the length of query protein is less than 120 residues, Rosetta [8] was called to generate 200 models. The 200 models were clustered by Rosetta and the centroid models of several large clusters were chosen as predicted *ab initio* models. During CASP8, Rosetta was used to generate models for several hard targets.

Model Ranking. All the models were assessed by our model quality assessment tool ModelEvaluator [9]. ModelEvaluator compared the secondary structure, solvent accessibility, contact map and beta-sheet topology of a model with those predicted from its primary sequence by the SCRATCH suite [10]. The comparison resulted in a number of features. The features were fed into Support Vector Machines to predict the GDT-TS score of the model. The predicted GDT-TS [11] scores were used to rank the models. The top five ranked models were submitted to CASP.

According to our preliminary assessments, MULTICOM-CLUSTER worked well on both easy and hard targets. It was very effective on high-accuracy targets. Particularly, it generated the models with the highest GDT-TS scores for a number of targets (T410, T0418, T0426, T0442, T0453, T0460, T0490_2).

2. Model Quality Assessment (QA)

In the QA category, MULTICOM-CLUSTER used a novel *hybrid* approach to assess both global and local model quality of CASP8 server models. It first used ModelEvaluator to predict the GDT-TS score for each model. The models were ranked by predicted GDT-TS scores. The top five ranked models were chosen as reference models. Then each model was superimposed against the top five models one by one using the structure comparison tool TM-Score [12], which resulted in a GDT-TS score. The average GDT-TS score between each model and the five reference models was the predicted global quality of the model. This method is a hybrid combination of the single-model evaluation method and the model comparison approach. We preliminarily evaluated the method on 113 targets whose experimental structures had been released. We computed the global correlation between true GDT-TS scores and predicted GDT-TS scores for all the models, and the per-target correlation for the models of each target. The global correlation and the average per-target correlation are 0.92 and 0.90 respectively. The loss, the difference of the GDT-TS score between the best model with highest real GDT-TS score and the No. 1 model ranked by the predicted GDT-TS score, was calculated. The average loss of GDT-TS score on 113 targets is 6.0. The method was more accurate than single-model approaches and achieved the performance comparable to the clustering or consensus QA methods.

The hybrid method was also used to predict the local quality of a residue in a model. During the structure comparison between a model and each reference model, the superimposition of the model and the reference model was generated. The distance between the position of a residue in the model and its counterpart in the reference model was calculated. The average distance over the five reference models was calculated and used as the predicted local quality of the residue.

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. Soeding J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics*. **21**, 951-960.
3. Sadreyev R.I., Grishin N.V. (2004). Quality of alignment comparison by COMPASS improves with inclusion of diverse confident homologs. *Bioinformatics*. **20**, 818-828.
4. Zhou H., Zhou Y. (2005). SPEM: Improving multiple-sequence alignment with sequence profiles and predicted secondary structure. *Bioinformatics*. **21**, 3615-3621.
5. Cheng J., P. Baldi. (2006). A machine learning information retrieval approach to protein fold recognition. *Bioinformatics*. **22**, 1456-1463.
6. Cheng J. (2008). A multi-template combination algorithm for protein comparative modeling. *BMC Structural Biology*. **8**:18.
7. Sali A., Blundell T.L. (1993). Comparative protein modeling by satisfaction of spatial restraints. *J. Mol. Bio.* **234**, 779-815.
8. 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**, 209-25.
9. Wang Z., Tegge A.N., Cheng J. (2008). Evaluating the absolute quality of a single protein model using support vector machines and structural features. *Proteins*, in press.
10. Cheng J., Randall A., Sweredoski M., Baldi P. (2005). SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Research*. **33**, w72-76.
11. Zemla A. (2003). LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Research*. **31**, 3370-3374.

12. Zhang Y., Skolnick J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins*. **57**, 702-710.

MULTICOM-CMFR

Prediction of Tertiary Structure, Model Quality, Domain Boundary, Contact Map and Disorder Regions by MULTICOM-CMFR

J. Cheng[1,2], Z. Wang[1] and A. N. Tegge[2]
[1] *Computer Science Department,* [2] *Informatics Institute, University of Missouri, Columbia, MO 65211, USA*
chengji@missouri.edu

MUTICOM-CMFR is a server in the MULTICOM series. MULTICOM-CMFR participated in tertiary structure prediction, domain prediction, model quality assessment, disorder prediction and contact map prediction. Here we describe the methods in each category.

1. Tertiary Structure Prediction (TS) by MULTICOM-CMFR

The special features of both MULTICOM-CMFR include two-level template identification, alternative alignments, multi-template combination, and model evaluation. MULTICOM-CMFR and MULTICOM-RANK mainly differs in the level-1 template identification. The entire structure prediction process is described as follows. **(1) Level-1 Template Identification for Easy Targets**. A query protein was first searched against the NCBI Non-Redundant protein sequence database to identify homologous sequences by PSI-BLAST [1]. The group of homologous sequences was used to build a Position Specific Scoring Matrix profile (PSSM) for MULTICOM-CMFR. The PSSM profile was searched against the template protein sequence database by PSI-BLAST. The templates found by PSI-BLAST and their alignments were collected. The query-template alignments were ranked by e-values. **(2) Multi-Template Combination and Model Generation**. The pairwise query-template alignments produced in step (1) were combined into multiple sequence alignments to generate models [3]. (See MULTICOM-RANK abstract for details) **(3) Level-2 Template Identification for Hard Targets**. If less than five significant templates were found, a sensitive machine learning fold recognition method [4] was used to rank templates in a large template library. (See MULTICOM-RANK abstract for details) **(4) Alternative Alignments, Model Generation and Evaluation**. Five alternative alignment tools (MUSCLE [5], Lobster [6], SPEM [7], COMPASS [8], hhsearch) were used to generate alignments between the query and the templates identified in step (3). The alignments were fed into Modeller [9] to generate models. The models were evaluated by ModelEvaluator [10], and the top ranked models were submitted. (See MULTICOM-RANK abstract for details).

2. Model Quality Assessment (QA)

In the QA category, MULTICOM-CMFR is a single-model, structure-based model quality assessment method. MULTICOM-CMFR predicted the absolute quality score of a single protein model from its structural features as in ModelEvaluator [10].

Given a model, MULTICOM-CMFR compared the secondary structure, solvent accessibility, beta-sheet topology and contact map extracted from the model with those predicted from the primary sequence by the SCRATCH suite. The comparison resulted in a number of fitness scores such as secondary structure matching scores. Since sequence-based predictions are reasonably good, the structural features of a good model are expected to match better with the sequence-based predicted features than those from a bad model. So the fitness scores are informative indicators of the model quality. These fitness scores were fed into a support vector machine trained on a sub set of CASP6 models, which predicted the quality score (i.e. GDT-TS) of the model. The methodology of MULTICOM-CMFR is similar to MULTICOM-REFINE (QA) except that MULTICOM-CMFR discarded partial models in which the coordinates of some residues are missing.

We preliminarily assessed the methods based on the released experimental structures of 113 targets. We calculated the GDT-TS scores of CASP8 models based on experimental structures. We computed the global correlation between true GDT-TS scores and predicted GDT-TS scores for all the models, and the per-target correlation of the models of each target. The global correlation is 0.78 and the average per-target correlation is 0.75. The average loss, the average difference of the GDT-TS score between the best model with the highest real GDT-TS score and the No. 1 models ranked by predicted GDT-TS scores, was 7.05. According to the results, MULTICOM-CMFR seemed to work pretty well among single-model, non-meta methods.

3. Hybrid Domain Boundary Prediction (DP) by MULTICOM-CMFR

MULTICOM-CMFR used a hybrid approach to predict protein domain boundaries. If a good model based on one or more significant templates was generated for a query protein, the structure-based domain parser PDP [11] was called to parse the models into domains similarly as in DOMAC [12]. If no good model was generated for a query protein, a novel *ab intio* domain prediction method (manuscript in preparation) was called to predict domain boundaries. The method first searched query protein against the NCBI non-redundant protein sequence database in order to identify homologous protein sequences using PSI-BLAST. The pairwise alignments between the query and each homologous protein sequence were used to identify *evolutionary signals* as candidate domain boundaries. A window of sequence and structural features including sequence profile, predicted secondary structure and solvent accessibility around the candidate domain boundaries were extracted. These features were fed into a pre-trained support vector machine to predict if the candidate site was a domain boundary. The predicted domain boundary sites were used to cut proteins into domains.

4. **Contact Map Prediction (RR)**

MULTICOM-CMFR used a re-trained 2-Dimensional Recursive Neural Network [13] to predict residue-residue contacts.

5. Disorder Region Prediction (DR)

Disorder prediction was first predicted by a 1-Dimensional Recursive Neural Network (1D-RNN) taking as input the profile of the sequence, and predicted secondary structure and solvent accessibility [15]. The predicted disorder probabilities of the residues were re-scaled so that the ratio of residues with disorder probability >= 0.5 is close to the ratio of the disorder residues in the training dataset used to train 1D RNN [16].

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. Soeding J. (2005). Protein homology detection by HMM-HMM comparison. Bioinformatics. 21, 951-960.
3. Cheng J. (2008). A multi-template combination algorithm for protein comparative modeling. BMC Structural Biology. 8:18.
4. Cheng J., P. Baldi. (2006). A machine learning information retrieval approach to protein fold recognition. Bioinformatics. 22, 1456-1463.
5. Edgar R.C. (2004). MUSCLE: multiple sequence alignment with accuracy and high throughput. Nucleic Acids Research. 32, 1792-97.
6. Edgar R.C., Sjolander K. (2003). SATCHMO: sequence alignment and tree construction using hidden Markov models. Bioinformatics. 19, 1404-1411.
7. Zhou H., Zhou Y. (2005). SPEM: Improving multiple-sequence alignment with sequence profiles and predicted secondary structure. Bioinformatics. 21, 3615-3621.
8. Sadreyev R.I., Grishin N.V. (2004). Quality of alignment comparison by COMPASS improves with inclusion of diverse confident homologs. Bioinformatics. 20, 818-828.
9. Sali A., Blundell T.L. (1993). Comparative protein modeling by satisfaction of spatial restraints. J. Mol. Bio. 234, 779-815.
10. Wang Z., Tegge A.N., Cheng J. (2008). Evaluating the absolute quality of a single protein model using support vector machines and structural features. Proteins, in press.
11. Alexandrov N, Shindyalov I. (2003). PDP: protein domain parser. Bioinformatics. 19, 429–430.
12. Cheng J. (2007). DOMAC: An accurate, hybrid protein domain prediction server. Nucleic Acids Research. 35, w354-w356.
13. Cheng J., Baldi P. (2005) Three-stage prediction of protein beta-sheets by neural networks, alignments, and graph algorithms. Bioinformatics, 21(Suppl 1), i75-84.
14. Cheng J., Baldi P. (2007). Improved residue contact prediction using support vector machines and a large feature set. BMC Bioinformatics. 8, 113, 2007.
15. Cheng J., Sweredoski M., Baldi P. (2005). Accurate Prediction of Protein Disordered Regions by Mining Protein Structure Data, Data Mining and Knowledge Discovery. 11, 213-222.
16. Hecker J., Yang J., Cheng J. (2008). Protein Disorder Prediction at Multiple Levels of Sensitivity and Specificity. BMC Genomics. 9, (S1):S9.

Multi-Template Combination, Alternative Alignments, Model Evaluation and Contact Predictions by MULTICOM-RANK

J. Cheng[1,2], A. N. Tegge[2] and Z. Wang[1]
[1] *Computer Science Department,* [2] *Informatics Institute, University of Missouri, Columbia, MO 65211, USA*
chengji@missouri.edu

MULTICOM-RANK is a server in the MULTICOM series. MULTICOM-RANK took part in tertiary structure prediction, model quality assessment, and contact map prediction.

1. **Tertiary Structure Prediction (TS)**

The special features of MULTICOM-RANK include two-level template identification, alternative alignments, multi-template combination, and model evaluation.

Level-1 Template Identification for Easy Targets. A query protein was first searched against the NCBI Non-Redundant protein sequence database to identify homologous sequences by PSI-BLAST [1]. The group of homologous sequences was used to build a hidden Markov model (HMM) profile. The HMM profile was searched against the template protein HMM profile database by hhsearch [2]. The templates found by hhsearch and their alignments were collected. The query-template alignments were ranked by e-values.

Multi-Template Combination and Model Generation. The most "significant" (i.e. e-value < -20 and cover ratio > 75%) query-template alignment was selected to combine with alignments ranked below it into a multiple sequence alignment by a greedy multi-template combination algorithm [3]. The process was repeated up to five times to generate up to five multiple sequence alignments, starting from each of the five top-ranked "*significant*" query-template alignments. Each multiple sequence alignment and the associated template structures were fed into Modeller [4] to generate 10 models, from which the model with minimum energy was selected as a predicted model. The models ordered by the e-values of the query-template alignments were submitted to CASP8. If five models could be generated in this way, the prediction process was done.

Level-2 Template Identification for Hard Targets. If less than five significant templates were found, a sensitive machine learning fold recognition method [5] was called to rank templates in a large template library. The top 50 templates were selected for further analysis.

Alternative Alignments, Model Generation and Evaluation. Five alternative alignment tools including MUSCLE [6], hhsearch, lobster [7], SPEM [8], and COMPASS [9] were used to generate *alternative* alignment for the query and each template. The alternative alignment approach may generate a better alignment than a single alignment tool. The 250 (50 × 5) query-template alignments were fed into Modeller to generate 250 models. The models were evaluated and ranked by our model evaluation tool ModelEvaluator [10]. The top ranked models are submitted to CASP. According to our preliminary assessments, MUTICOM-RANK worked well, especially on high-accuracy targets. MULTICOM-RANK produced the models with the highest GDT-TS scores for targets T0408, T0411, T0418, T0425, T0428, T0453, and T0509.

2. **Contact Map Prediction (RR)**

MULTICOM-RANK used support vector machines [11] to predict contact maps. Both methods are *ab initio* approaches without using any template information.

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. Soeding J. (2005). Protein homology detection by HMM-HMM comparison. Bioinformatics. 21, 951-960.
3. Cheng J. (2008). A multi-template combination algorithm for protein comparative modeling. BMC Structural Biology. 8:18.
4. Sali A., Blundell T.L. (1993). Comparative protein modeling by satisfaction of spatial restraints. J. Mol. Bio. 234, 779-815.
5. Cheng J., P. Baldi. (2006). A machine learning information retrieval approach to protein fold recognition. Bioinformatics. 22, 1456-1463.
6. Edgar R.C. (2004). MUSCLE: multiple sequence alignment with accuracy and high throughput. Nucleic Acids Research. 32, 1792-97.

7. Edgar R.C., Sjolander K. (2003). SATCHMO: sequence alignment and tree construction using hidden Markov models. Bioinformatics. 19, 1404-1411.
8. Zhou H., Zhou Y. (2005). SPEM: Improving multiple-sequence alignment with sequence profiles and predicted secondary structure. Bioinformatics. 21, 3615-3621.
9. Sadreyev R.I., Grishin N.V. (2004). Quality of alignment comparison by COMPASS improves with inclusion of diverse confident homologs. Bioinformatics. 20, 818-828.
10. Wang Z., Tegge A.N., Cheng J. (2008). Evaluating the absolute quality of a single protein model using support vector machines and structural features. Proteins, in press.
11. Cheng J., Baldi P. (2007). Improved residue contact prediction using support vector machines and a large feature set. BMC Bioinformatics. 8, 113, 2007.

MULTICOM-REFINE

Model Combination, Refinement and Assessment by MULTICOM-REFINE

J. Cheng[1,2], A. N. Tegge[2] and Z. Wang[1]
[1] *Computer Science Department,* [2] *Informatics Institute, University of Missouri, Columbia, MO 65211, USA*
chengji@missouri.edu

MULTICOM-REFINE is a server in our MULTICOM series. MULTICOM-REFINE participated in tertiary structure prediction and model quality assessment, which are described in the following two sections.

1. Tertiary Structure Prediction

MULTICOM-REFINE server focused on *ranking* and *refining* structural models generated by our three other MULTICOM servers: MULTICOM-CMFR, MULTICOM-RANK and MULTICOM-CLUSTER. It worked as follows.

Model Ranking. The models of a target generated by MULTICOM-CMFR, MULTICOM-RANK and MULTICOM-CLUSTER were collected together. ModelEvaluator [1] was used to predict the quality (GDT-TS [2] scores) of the models. The top 50% of the models generated by MULTICOM-CMFR and MULTICOM-RANK, in addition to all the models generated by MULTICOM-CLUSTER, were selected for model combination and refinement.

Model Combination and Refinement. All the selected models were pooled together and ranked by their predicted GDT-TS scores. The models were combined and refined by a structure alignment-based *global-local* model combination algorithm as follows. First, one model out of the top five ranked models was selected as an initial seed model. The model was compared against all other models using a structure-comparison tool TM-Score [3]. The models where 80% of the regions can align with the seed model with less than 4 Å RMSD were considered globally similar to the seed model and selected for combination. The seed model and the selected models were used as structure templates for Modeller 7v7 [4] in order to re-generate 10 models for the query protein. The model with the minimum Modeller energy was selected as a refined model for the target protein. By repeating the process up to five times, up to five refined models were generated from the top five ranked models. The *global* model combination procedure worked for easy targets where many similar models were generated.

If no globally similar models were found, which often happened for hard targets, a *local* model combination algorithm was used to combine a seed model with other locally similar models. The seed model was compared against other models using TM-Score. The long fragments of the models that can align with the seed model with RMSD < 3 Å and GDT-TS score > 50 were selected. The minimum length of the fragment was set to 80 residues initially. It was repeatedly reduced by 5 if no similar fragments were found during an iteration of model comparison. The structures for the fragments and the initial seed model were fed into Modeller to generate 10 models, and the model with minimum energy was chosen as a refined model. After the *global-local* combination algorithm was finished, the five refined models based on the top five initial seed models were submitted to CASP.

Preliminary Results. Our *global-local* model combination method is a kind of iterative modeling and refinement technique that combines models in order to improve model quality. It combines globally similar models directly and only integrates similar local structures from locally similar models. This approach performs very well on high-accuracy targets. According to our preliminary assessment, the refinement method improved the model quality for a number of targets when compared to the original models. It produced the models with the highest GDT-TS score for T0404, T0459, T0475, and T0506_2.

2. Model Quality Assessment (QA)

In the QA category, MULTICOM-REFINE is a single-model, structure-based model quality assessment method. MULTICOM-REFINE predicted the absolute quality score of a single protein model from its structural features as in ModelEvaluator [1].

Given a model, MULTICOM-REFINE compared the secondary structure, solvent accessibility, beta-sheet topology and contact map extracted from the model with that predicted from the primary sequence by the SCRATCH suite [5]. The comparison resulted in a number of fitness scores such as secondary structure matching scores. Since sequence-based predictions are reasonably good, the structural features of a good model expect to match better with the sequence-based predicted features than those from a bad model. So the fitness scores are informative indicators of the model quality. These fitness scores were fed into a support vector machine trained on CASP6 and CASP7 models to predict the quality score (i.e. GDT-TS) of the model. The basic idea of MULTICOM-REFINE (QA) is similar to MULTICOM-CMFR (QA) except that MULTICOM-REFINE can handle partial models in which the positions of some residues are missing.

We preliminarily assessed the methods based on the released experimental structures from 113 targets. We calculated the GDT-TS scores of CASP8 models based on experimental structures. We computed the global correlation between true GDT-TS scores and predicted GDT-TS scores for all the models, and the per-target correlation for the models of each target. The global correlation is 0.80 and the average per-target correlation is 0.73. The loss, the difference of the GDT-TS score between the best model with highest real GDT-TS score and the No. 1 model ranked by the predicted GDT-TS score, was calculated. The average loss of GDT-TS score on 113 targets is 8.0. According to the results, MULTICOM-REFINE seemed to work pretty well among single-model, non-meta methods.

1. Wang Z., Tegge A.N., Cheng J. (2008). Evaluating the absolute quality of a single protein model using support vector machines and structural features. *Proteins*, in press.
2. Zemla A. (2003). LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Research*. **31**, 3370-3374.
3. Zhang Y., Skolnick J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins*. **57**, 702-710.
4. Sali A., Blundell T.L. (1993). Comparative protein modeling by satisfaction of spatial restraints. *J. Mol. Bio.* **234**, 779-815.
5. Cheng J., Randall A., Sweredoski M., Baldi P. (2005). SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Research*. **33**, w72-76.

MUMSSP

Side chain modeling and loop refinement: Homology modeling of 5 CASP targets

M. R. Saberi[1], A. Baratian[1], M. Moussavi[2], M. Zabihi[2]

[1]-Medicinal Chemistry Department, School of Pharmacy, Mashhad University of Medical Sciences, Vakilabad Boulevard, Mashhad, 91775-1365, IRAN

[2]-Biology Department, Azad University, Rahnamaee Street, Mashhad, IRAN

saberimr@mums.ac.ir

When modeling proteins, all modelers go through usual procedures i.e. searching proper template(s), finding the best alignment(s), predicting the most accurate secondary structure prediction, forecast folding of the protein in super secondary structure and tertiary structure, qualify and assess all gathered data and finally do protein modeling and assessment. They usually go back and improve the models by using different template(s) and alignment(s) and repeat modeling until the best model fulfills them and meet the reality. Lots of sites, servers, computers and software are exploited during a protein modeling project however some modelers develop their own facilities including software and algorithms. The challenge appears when trying to resolve a model for the entire protein including loops. Loops are parts of proteins which fold as they want and can affect the quality and accuracy of a protein.

We report here a deep trial study of loop refinement in improvement of 7 models for 5 CASP targets. As mentioned, this study utilized usual procedures to fined template(s), alignment, secondary structure prediction, folding prediction, motif prediction, modeling and quality assessment of CASP targets. UCLA, NCBI and EBI sites, ExPasy, PDB, FUGUE and PSSM servers and many other bioinformatics web sites and servers as well as software such as MODELLER 9v3, MolMol, ViewerLite, Autodock, Chem3D, Rasmol, etc. applied for modeling the targets. What_Check, ERRAT, and verify3D were the methods of protein 3D structure assessment to assess stereochemistry, atom environment and solvent accessibility of

models respectively. Trial-error method was the choice until no more improvement was achieved for models. Then models were energy minimized as whole and improper loops separately. Different windows were selected on loops for energy minimization and the windows were shrunk until a few residues remained unrefined .Problematic residues in loops were then selected and minimized in third step until changing the conformation of those residues were not advantageous any more. In the forth step other residues of neighbor segments of the protein in a 3D environment which were not necessarily the neighbor residues in the raw sequence were minimized with the problematic loop residues together in a box using MODELLER's loop model class. Then, the whole proteins were energy minimized by Means of MM+. RMS gradient was decreased in a step wise approach. It was of surprise to see that energy minimization, while improving model's performance in tests dramatically, could damage the structure's performance if excessively applied. To avoid damage to 3D structure due to excessive refinement, a very conservative approach was selected in this step. Finally, side chains of all of protein's amino acids were selected and significantly energy minimized. Of course model improvement was tracked during the model refinement applying What_Check, ERRAT, and verify3D methods. 7 CASP models submitted by our team looks promising and show high quality compared to the released structures of the targets.

We think there is still way to set up satisfying method for enhancing the folding of loops due to the nature of loops, their exposure to the surface of proteins and their size. But when facing a protein in which loops could play a critical role like antibodies or proteins interacting other proteins one must always be careful about the quality of the loops.

1. Saberi M.R., Razazan A., Ramezani H. and Baratian A. How do the web facilities help predictors from head to toe of homology modeling? CASP6 Abstract book, P 166.
2. Saberi M. R., Baratian A., Sadeghian H. Loop refinement and geometry optimization: key steps in protein modeling. CASP7 Abstract book, P 79-80.
3. Cheng X, Cui G, Hornak V, Simmerling C. (2005) Modified replica exchange simulation methods for local structure refinement. J Phys Chem B Condens Matter Mater Surf Interfaces Biophys. Apr, 28;109(16):8220- 30.

MUProt

Model Ranking, Model Combination and Refinement by MUProt

J. Cheng[1,2], Z. Wang[1] and A. N. Tegge[2]

[1] *Computer Science Department,* [2] *Informatics Institute, University of Missouri, Columbia, MO 65211, USA*

chengji@missouri.edu

MUProt is a server in our MUTICOM series. Like the MULTICOM-REFINE server, MUprot server focused on *ranking* and *refining* models generated by our three other MULTICOM servers, including MULTICOM-CMFR, MULTICOM-RANK and MULTICOM-CLUSTER. MUProt differs from MULTICOM-REFINE mainly in the model ranking step.

Model Ranking. The models of a protein target generated by MULTICOM-CMFR, MULTICOM-RANK and MULTICOM-CLUSTER were collected together. The models were clustered by Spicker [1]. The model closest to the centroid of the largest cluster was ranked first. ModelEvaluator [2] was used to predict the quality (GDT-TS scores) of the remaining models. All the remaining models were ranked by their predicted GDT-TS [3] scores. The top five models were selected as reference models. All the models were compared with each of the reference models by TM-Score [4]. TM-Score produced a GDT-TS score for each comparison. The average GDT-TS score over the five reference models was used as the predicted quality of each model. The models were re-ranked by the predicted quality score.

Model Combination and Refinement. The top five ranked models were used as the seed model to combine with other models by a *global-local* model combination procedure. This step is the same as in MULTICOM-REFINE (see MULTICOM-REFINE's abstract for details). The five refined models resulted from model combination were submitted to CASP.

According to the preliminary assessments, MUProt's ranking and model refinement approach improved the quality of original models in many cases. For instance, it generated models with highest GDT-TS scores for targets T0390, T0404, T0426, and T0432.

In addition to tertiary structure prediction, MUProt also took part in residue-residue contact prediction. It used a support vector machine and a large feature set to predict contact maps [5]. The method is similar to MULTICOM-RANK except that it was trained with a few more input features.

1. Zhang Y., Skolnick J. (2004). SPICKER: a clustering approach to identify near-native protein folds. *J. Comput. Chem*. **25**, 865-871.
2. Wang Z., Tegge A.N., Cheng J. (2008). Evaluating the absolute quality of a single protein model using support vector machines and structural features. *Proteins*, in press.
3. Zemla A. (2003). LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Research*. **31**, 3370-3374.
4. Zhang Y., Skolnick J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins*. **57**, 702-710.
5. Cheng J., Baldi P. (2007). Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinformatics*. **8**, 113, 2007.

MUSTER

MUSTER: a single-threading server using sequence and structure profile-profile alignment and multiple template libraries

S. Wu and Y. Zhang
Center for Bioinformatics and Department of Molecular Bioscience, University of Kansas, 2030 Becker Dr, Lawrence, KS 66047
yzhang@ku.edu

MUSTER[1] is a recently developed threading program aiming at improving the accuracy of the sequence profile-profile alignment algorithm with the help of structural profile information. Besides the sequence profile-profile alignment, the alignment score of MUSTER incorporates five types of structural information: (1) match of secondary structures of query and templates; (2) alignment of sequence-based query profile with structured-based template profile; (3) match of solvent accessibility of query and templates; (4) match of torsion angles (φ and ψ) between query and templates; (5) hydrophobic scoring matrix. In a benchmark test of 500 non-homologous proteins, it was found that the average TM-score[2] of the first threading alignment to native is nearly 5% higher than the PPA algorithm[3] which is based on sequence profile-profile alignment and secondary structure match.

The template library of MUSTER is collected from the PDB[4] with homologous proteins with sequence identity >70% removed. However, we found that the template library thus constructed often missed better template structures for some targets. To have a complete pool of the template structures, we created 7 non-redundant libraries with sequence identity cutoffs <70%. We first construct the 1st library as usual from all the solved structures in PDB. For multiple homologous proteins, we select the one which was solved in the highest resolution. For multiple-domain proteins, we keep both the full-chain and the individual domains in the library. To construct the 2nd library, we first collect non-redundant proteins from the PDB structures which are exclusive to the 1st library. Then we add the proteins with missed folds (based on sequence identity) from the 1st library with the purpose of keeping the 2th library complete. For creating the 3rd library, we first select non-homologous proteins from those which are not in the 1st and the 2nd libraries and then add the proteins with missed folds from the 1st and the 2nd libraries to make the 3rd library complete. This procedure is repeated 7 times until seven different complete libraries (the number of 7 is picked up somewhat arbitrarily) are generated. For each target, we thread the sequence through the 7 template libraries independently and rank all the alignments together by their Z-scores, and the redundant threading alignments from the same template are removed at the same time. In this way, the templates with correct folds missed in the 1st library may be found by MUSTER on other libraries, which will eventually increases the likelihood of MUSTER to identify correct folds.

After the threading, all the alignments are clustered based on structural similarity TM-score[2] >0.9/0.85/0.8 for Easy/Medium/Hard targets, respectively. Following the clustering, all threading alignments in the same cluster are fed into MODELLER[5] to build a full-length model where the spatial restraints represented by C-alpha distances collected from the clusters are used as well. We submitted top five models ranked by the highest Z-score of threading alignments in the clusters.

1. Wu, S. & Zhang, Y. (2008). MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. *Proteins* **72**, 547-556.
2. Zhang, Y. & Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins-Structure Function and Bioinformatics* **57**, 702-710.
3. Wu, S. & Zhang, Y. (2007). LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic acids research* **35**, 3375-3382.

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. Sali, A. & Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **234**, 779-815.

nFOLD3

Fully automated protein fold recognition using nFOLD3

L.J. McGuffin
School of Biological Sciences, University of Reading, Whiteknights, Reading RG6 6AS, UK
l.j.mcguffin@reading.ac.uk

Tertiary structure predictions were submitted in the automatic server category using the latest version of the nFOLD protocol. The previous versions of nFOLD[1,2] aimed to improve mGenTHREADER[2,3] through the incorporation of several additional model quality assessment scores as inputs to the underlying neural network. These extra inputs included the Secondary Structure Element Alignment (SSEA) score and model quality assessment scores from MODCHECK[4] and ProQ[5].

The latest version of the method, nFOLD3, maintained the original idea, in that it attempted to select the optimum models using a consensus of model quality assessment programs (MQAPS). However, the ModFOLD[6,7] model quality assessment program was used to rank models, which were built using alignments from several alternative profile-profile methods.

The SP3[8], SPARKS[8], and HHsearch[9] methods were run in house for each target against a bespoke template library. A series of profile-profile alignments were generated and 3D models were built using Modeller[10]. Each model was then assessed individually using ModFOLD version 1.1. The ModFOLD output score for each model was then combined with the initial rank of the target-template alignment assigned by each individual alignment method.

Preliminary results indicated that using ModFOLD for model selection improved the performance compared with using the rankings from each individual alignment method alone. Overall, the nFOLD3 method appears to be competitive with several other popular independent servers. The nFOLD3 server also appears to have predicted the best model for CASP8 target T0417, compared with those predicted by all other servers.

The nFOLD3 web server is available at the following URL:
http://www.reading.ac.uk/bioinf/nFOLD/

1. Bryson,K., McGuffin,L.J., Marsden,R.L., Ward,J.J., Sodhi,J.S. & Jones,D.T. (2005) Protein Structure Prediction Servers at University College London. *Nucleic Acids Res.* **33**, W36-38.
2. Jones,D.T., Bryson,K., Coleman,A., McGuffin,L.J., Sadowski,M.I., Sodhi,J.S. & Ward,J.J. (2005) Prediction of novel and analogous folds using fragment assembly and fold recognition. *Proteins*. **61** (S7), 143-151.
3. McGuffin,L.J. & Jones,D.T. (2003) Improvement of the GenTHREADER method for genomic fold recognition. *Bioinformatics.* **19**, 874-881.
4. Pettitt,C.S., McGuffin,L.J. & Jones,D.T. (2005) Improving sequenced based fold recognition by use of 3D model quality assessment. *Bioinformatics.* **21**, 3509-3515.
5. Wallner,B. & Elofsson,A. (2003) Can correct protein models be identified? *Protein Sci*. **12**, 1073-1086.
6. McGuffin,L.J. (2008) The ModFOLD Server for the Quality Assessment of Protein Structural Models. *Bioinformatics.* **24**, 586-587.
7. McGuffin,L.J. (2007) Benchmarking consensus model quality assessment for protein fold recognition, *BMC Bioinformatics.* **8**, 345.
8. Zhou,H. & Zhou,Y. (2005) SPARKS 2 and SP3 servers in CASP6. *Proteins*. **61** (S7), 152-156.
9. Söding, J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*. **21**, 951-96.
10. Sali,A. & Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol*. **234**, 779-815.

Using the CASP8 experiment in undergraduate course on structure prediction

M. Schushan[1] and N. Ben-Tal[1]
[1]*Department of Biochemistry, George S. Wise Faculty of Life Sciences, Tel-Aviv University 69978 Tel-Aviv, Israel*
mayaschu@tauex.tau.ac.il

We initiated a course aimed at structure prediction for undergraduate students. The course was offered within the Faculty of Life Science and was open for all students but most of the participants were from the Bioinformatics track at Tel-Aviv University, Israel. The course focused on combining homology modeling with other publicly available sequence and structure prediction tools to produce model-structures. The course subjects introduced various well-known bioinformatical methods that can be utilized for protein modeling. This included methods for sequence homology detection, construction of multiple sequence alignments, fold recognition approaches, secondary structure prediction methods and also structure-based function prediction approaches. Since homology modeling was the main course subject, we gave large emphasis to proper template detection and selection. We also accentuated methods that enable the generation of accurate query-template pairwise alignments as a crucial step for producing high quality models.

The final task for the course was participating in the CASP8 experiment. Using the tools presented in the course, the students incorporated a combined computational modeling approach, consisting of homology modeling, fold recognition, secondary structure prediction and evolutionary conservation analysis. In brief, templates were detected using sequence homology tools and fold recognition methods. Query-template pairwise alignments were generated using multiple sequence alignments and fold recognition methods. The alignments were then refined using structure prediction algorithms. Models were built via the NEST homology modeling algorithm[45]. Model evaluation and additional pairwise alignment refinement were conducted using the ConSurf webserver ([2], http://consurf.tau.ac.il). Overall, the course students submitted ten models for the CASP8 experiment. The students appeared to be very excited about the course and in particular about the final project.

The submitted models showed the significance of using evolutionary conservation analysis for guiding model evaluation and, to some extent, for model refinement. Generally speaking, it is well known that protein structures present an evolutionary conserved core while non-functional solvent-accessible residues are variable[3]. This expected evolutionary conservation pattern can thus be utilized to examine the reliability of a produced model-structure. Therefore, evolutionary conservation scores were mapped on the ten generated models using the ConSurf webserver ([46], http://consurf.tau.ac.il) in order to evaluate the models' validity. Moreover, in cases which the evolutionary conservation patterns of the model-structures did not match the expected distributions, we reassessed both the template selection and the pairwise alignments to improve the initial models.

1. 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., & Honig, B. (2003). Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling. *Proteins* **53**, 430–435.
2. Landau, M., Mayrose, I., Rosenberg, Y., Glaser, F., Martz, E., Pupko, T., & Ben-Tal, N. (2005). ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res*. **33**, W299–W302.
3. Branden, C. & Tooze, J. (1999). *Introduction to Protein Structure*, 2nd edit., Garland Publishing, Inc., NewYork.

Information-based alignments, local dominating set refinement and secure implementation in a server.

R.W. Harrison[1],
[1] – *Department of Computer Science Georgia State Unviersity*
rwh@gsu.edu

Panther_server is a single alignment – single model server using a profile-profile alignment tool and a molecular mechanics modeling protocol. It selected up to five top sequence alignments and could produce up to five different submissions. The ranking between targets was solely based on sequence alignment quality, and no model quality ranking was used. Thus, model 1 was not always the "best" model, but generally, at least one of the five models was of competitive quality for comparative modeling.

The sequence alignment used a full-dynamic programming algorithm where the cost function was based on the Kullback entropy (1) between the query and database sequence profiles. Profiles were calculated with psi-blast[1] and corrected for sample bias using the equilibrium frequency distribution of amino acids. No gap penalty was used or needed.

$$(1)\ KE = \log \frac{\left(\sum p_{query}\, p_{database}\right)^2}{\sum p^2{}_{query} \sum p^2{}_{database}}$$

Since full dynamic programming is computationally expensive a quicker prescreen algorithm simply searched for consecutive maxima of the Kullback entropy, and reported possible hits that were searched with the full algorithm. The Kullback entropy improves the signal to noise especially when compared with measures like the correlation between the profiles, and makes the alignment clearer. It does not correct for errors in profile estimation.

Model refinement was done with the program AMMP[2] using the current potential set (version tuna). In order to build insert regions, a database of common 10-mer protein structures were overlapped with the ends of the insert and the closest fit was used as a basis for distance restraints for the peptide backbone. Insertions that were too long for reliable identification of a 10-mer were truncated.

Side chains were refined using an annealing procedure based on local dominating sets. The local dominating set consists of a partition of the contact graph between side chains, where the total depth along any branch of the graph is limited. It can be loosely thought of as the side chain and its neighbors. Members of the local dominating set were given small random deviations in the c_α c_β torsion. The model was then minimized and if the energy decreased (or randomly if the energy increased), the changes were accepted.

The server was designed to be secure and to achieve a good load balance in a distributed computing environment. Input data from the website (http://bmcc3.cs.gsu.edu) were sanitized and written into an xml description of the task to be accomplished. This makes it difficult to implement a cross-site scripting attack because the potential attacker does not directly input commands. The xml packet was sent to an internal server (via the loopback interface 127.0.0.1) which parsed it and scheduled the tasks based on system load. Each stage parses the xml to ensure that it is correctly formed, and the xml describes the parameters and data for the calculation, but does not give commands directly to the server.

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. Harrison, R.W., Chatterjee D., & Weber I.T. "Analysis of six protein structures predicted by comparative modeling techniques." (1995) Proteins: Structure Function and Genetics 23:463-471

Protein structure prediction using a probabilistic model of local structure

W. Boomsma[1], M. Borg[1], J. Frellsen[1], T. Harder[1], K. Stovgaard[1], J. Ferkinghoff-Borg[2], A. Krogh[1], K.V. Mardia[3] and T. Hamelryck[1]

[1] - Department of Biology, University of Copenhagen,

[2] - Ørsted-DTU, Technical University of Denmark,

[3] - Department of Statistics, University of Leeds

thamelry@binf.ku.dk

We recently developed a generative probabilistic model of local protein structure, which can be considered as a potential alternative to the popular fragment assembly method. CASP8 was an opportunity to test the sampling performance of this new method. Our long-term goal is a complete description of the structure prediction problem in terms of probabilistic models. However, the modeling of non-local interactions is still in very early development, and we had limited expectations for our overall performance in the current CASP exercise.

Sampling procedure

A Markov Chain Monte Carlo (MCMC) simulation procedure was implemented using the generalized multi-histogram method[1]. As a proposal distribution, we used a probabilistic model of the local structural preferences of a protein backbone (TorusDBN[2]). The model uses a bivariate angular distribution to capture the dihedral bond angle preferences in continuous space, and makes it possible to sample candidate protein conformations that are locally compatible with a given amino acid sequence, or to resample parts of a structure while maintaining consistency along the chain. The produced structures have high quality local structure, and in contrast to fragment assembly based methods, TorusDBN is a well-defined probability distribution, making it a more natural component of an MCMC simulation. As input to the local model we used the predicted secondary structure labeling from PSIPRED[3].

Energy function

In addition to the energy described by the local model, we used three energy terms, describing hydrogen-bonding, compactness and multi-body contacts, respectively. The first was based on a energy term known from the literature[4], while the latter two are very preliminary. Currently, the non-local energy terms are the greatest weakness of our method. We hope to make significant progress in this area in time for the next CASP.

Clustering procedure

A new clustering technique was designed to handle the large number of samples produced by the simulation procedure. Proteins were represented using a space-curve representation, where each conformation can be described as a 30-dimensional vector[5]. This representation made it possible to cluster protein sets with millions of structures, using the Euclidean distance between these vectors as a measure of structural deviation. The 10% lowest energy structures produced during sampling were clustered using this approach.

Structure refinement

Both the lowest energy structures obtained during simulation and structures from the largest clusters were investigated in greater detail. Side-chains were added using IRECS[6], and the CHARMM program[7] with an implicit solvent forcefield was used for refinement.

Results

We submitted 5 predictions for assessment. For two of these, our predictions had a similar fold, with a GDT_{TS} of ~0.5. Of the remaining three targets, two turned out to be largely disordered. For the last target we did not predict any native-like conformations, which is partly due to using the wrong secondary structure as input to our model.

1. Ferkinghoff-Borg ,J. (2002) Optimized Monte Carlo analysis for generalized ensembles. Eur. Phys. J. B. 29, 481-484.
2. Boomsma,W., Mardia,K.V., Taylor,C.C., Ferkinghoff-Borg,J., Krogh,A. & Hamelryck,T. (2008). A generative, probabilistic model of local protein structure. Proc. Natl. Acad. Sci. U.S.A., 105, 8932-8937.

3. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292, 195-202.
4. Fabiola,F., Bertram,R., Korostelev,A. & Chapman,M.S. (2002). An improved hydrogen bond potential: Impact on medium resolution protein structures. Protein Sci. 11, 1415-1423.
5. Røgen,P. & Fain,B. (2003). Automatic classification of protein structure by using Gauss integrals. Proc. Natl. Acad. Sci. U.S.A., 100, 119-124.
6. Hartmann,C., Antes,I. & Lengauer,T. (2007). IRECS: A new algorithm for the selection of most probable ensembles of side-chain conformations in protein models. Protein Sci. 16, 1294-1307.
7. 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. Comput. Chem. 4, 187-217.

POEM
POEMQA

Performance of an all-atom free-energy approach for protein structure prediction and quality assessment

T. Strunk[1], F. Hoffgaard[2], K. Klenin[1] and W. Wenzel[1]
[1] - *Research Center Karlsruhe, Institute for Nanotechnology, Germany,*
[2] - *Technische Universität Darmstadt, FB Biologie, Germany*
Timo.Strunk@int.fzk.de , http://www.fzk.de/biostruct

De novo prediction of protein tertiary structure on the basis of amino acid sequence remains one of the outstanding problems in biophysical chemistry. We have developed an all-atom free energy forcefield PFF01/02[1] which stabilizes a wide array of proteins. Recently we have implemented these techniques in POEM@HOME (http://boinc.fzk.de), a world-wide distributed computational architecture.

We will discuss advantages and limitations of this approach for protein folding and structure prediction. Using this framework we have participated in CASP8 as human predictors in the context of quality assessment (POEM-QA) and structure prediction (POEM). In the structure prediction approach, decoy libraries are generated which we subsequently ranked in the refinement simulations using POEM@HOME. The structure corresponding to the lowest-energy cluster of the ranked decoys is then used for a prediction. We will discuss both the impact of decoy generation and of the refinement protocol on the quality of the prediction.

PFF02 enabled us to separate inadequately built decoys from near-native conformations in the POEM group. The universality of a free-energy approach like PFF02 allowed the usage of several structure generation techniques in a combined framework.

In the quality assessment approach, we ranked all submitted server models in short refinement simulations and find excellent selectivity: POEM-QA is capable, on the basis of its energy criterion alone, to select one of the best conformations from all submitted models.

1. Herges, T. & Wenzel, W. (2004) An all-atom forcefield for tertiary structure prediction of helical proteins. *Biophys. J.* 87, 3100.

pro-sp3-TASSER

Protein structure prediction by pro-sp3-TASSER

H. Zhou and J. Skolnick
Center for the Study of Systems Biology, School of Biology
Georgia Institute of Technology, 250 14th Street, N.W., Atlanta, GA 30318
skolnick@gatech.edu

Pro-sp3-TASSER is an automated protein structure prediction approach that uses the PRO-SP3 threading method to identify templates and TASSER[1] to refine the models. PRO-SP3 threading consists of five different threading scores, one derived from the SP3 [2] and four from PROSPECTOR_3[3] threading methods. Targets are classified by their SP3 threading Z-score into Easy, Medium and Hard categories. For Medium/Hard targets, alternative alignments are generated by a parametric approach[4] and good alignments selected by TASSER-QA[5]. The top templates identified by each threading score along with their alternative alignments are combined to derive contact and distant restraints for model refinement by short TASSER simulations. For Medium/Hard targets, chunk-TASSER [6] is also used to generate full length models. Multiple short TASSER or chunk-TASSER runs are used to generate an ensemble that has up to 150 full-length models. Subsequently, the top 20 models are selected from the ensemble by TASSER-QA. These are used to generate contact and distance restraints for longer TASSER modeling. Special attention is paid to possible multiple domain targets. We check the coverage of the top template as identified by its SP3 score; if more than 50 residues are unaligned, the unaligned and aligned regions are modeled separately in addition to modeling the full length target sequence. The separately modeled, possible domains are then overlapped onto the full length models in the second round of TASSER refinement. Other special cases are when the Z-score of the first SP3 template is 2.0 units higher than the second template or when a single template has > 50% sequence identity to the target; then, only models from the first or the single high sequence identity template are used in TASSER simulations. Final models are selected from both rounds of TASSER runs by TASSER-QA. Main-chain and side-chain atoms are rebuilt by PULCHRA[7] from the Cα only models.

1. Zhang, Y. and J. Skolnick, Automated structure prediction of weakly homologous proteins on genomic scale. Proc. Natl. Acad. Sci. (USA), 2004. 101: p. 7594--7599.
2. Zhou, H. and Y. Zhou, Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins, 2005. 58 p. 321--328.
3. Skolnick, J., D. Kihara, and Y. Zhang, Development and large scale benchmark testing of the PROSPECTOR 3.0 threading algorithm. Proteins, 2004. 56: p. 502--518.
4. Chivian, D. and D. Baker, Homology modeling using parametric alignment ensemble generation with consensus and energy-based model selection. Nucl. Aci. Res., 2006. 34: p. e112.
5. Zhou, H. and J. Skolnick, Protein model quality assessment prediction by combining fragment comparisons and a consensus Cα contact potential. Proteins, 2007. 71: p. 1211--1218.
6. Zhou, H. and J. Skolnick, Ab initio protein structure prediction using chunk-TASSER. Biophys. J., 2007. 93: p. 1510--1518.
7. Rotkiewicz, P. and J. Skolnick, Fast procedure for reconstruction of full-atom protein models from reduced representations. Journal of Computational Chemistry, 2008. 29: p. 1460--1465.

ProtAng

Graph clustering approach for domains delineation problem in protein structures

M. Milostan[1], P. Lukasiak [1,2] and J. Blazewicz [1,2]
[1] *– Poznan University of Technology, Institute of Computing Science, Poznan, Poland*
[2] *– Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland*
Maciej.Milostan@cs.put.poznan.pl

Structural domains [1] of a protein are regions that are either compact, globular modules, or are clearly distinguished from flanking regions. Domains can be viewed as semi-independent three-dimensional units in proteins; they may fold independently and may constitute units of evolution. These units are stabilized by various kinds of interactions or forces occurring among composing them amino acids. Among examples of such forces one can distinguish for example chemical bonds or compact amino acid packing enforced by solvent. The basic idea is to recognize, somehow, these interactions or spatial contacts and recognize domains on that basis. It is worth to note that such contacts could also appear between separate domains.

That fact makes problem harder to solve. In the following sections novel algorithm for decomposing tertiary structure into domains has been outlined.

Although it is hard to define domain as a formal entity, it is possible to provide
some basic features of the valid domain. A domain should have at least 40 residues, be compact, have small cross-domain interface and not too many segments. Segment is a fragment of sequence composing part of a domain (c.f. [2] and [3]).

The most straightforward approach to tackle the problem is to represent protein structure as a graph of contacts and then to partition the graph into stable clusters. In such a case one has to identify contacts and then convert each residue in protein chain into vertex in the graph, and represent each contact as an edge. For purpose of contact identification we used distances between geometrical centers of side-chains and distances between C_{Alpha} carbons.
Given the protein graph one can apply graph clustering approaches for determination of potential domains. However most efficient clustering methods need prior knowledge about number of clusters. To overcome this problem the idea of the structure coloring using simple rules has been proposed. The proposed method contains following steps: contact graph generation, identification of small stable substructures, merging these substructures into clusters and final refinement of the assignments.

Exemplary results showed that proposed method has large potential. For the test set presented in [3] and [4] the algorithm gives similar results to one of the other compared approaches or SCOP [5, 6] database.

The proposed method produces comparable results, in sense of assignment conformity, to the other approaches known from literature, however it has lower complexity. Therefore it can be useful in protein structures analysis.

1. Taylor WR (1999). Protein structural domain identification. *Protein Eng.* **12(3)**, 203-16.
2. Holm L. and C. Sander C. (1994). Parser for protein folding units. *Proteins: Structure, Function, and Genetics,* **19**, 256–268.
3. Xu Y. and Xu D. and Gabow HN (2000). Protein domain decomposition using a graph theoretic approach. *Bioinformatics*, **16**,12:1091-1104, 2000.
4. S. Jones, M. Stewart, A. Michie, M.B. Swindels, C. Orengo and J.M. Thornton Domain assignments for protein structures using a consensus approach: characterization and analysis. Protein Sci., 7:233–242, 1998.
5. A.G. Murzin, S.E. Brenner, T. Hubbard and C. Chothia SCOP: a structural classification of proteins database for the investigation of sequences and structures. Journal of Molecular Biology., 247:536–540, 1995.
6. A. Andreeva, D. Howorth, S.E. Brenner, T.J.P. Hubbard, C. Chothia and
   A.G. Murzin SCOP database in 2004: refinements integrate structure and sequence family data.
   Nucleic Acid Research., 32:226–229, 2004

ProteinShop

Protein Structure Prediction Using ProteinShop

Ch. Hu[1], N. Max[1,2] and S. Crivelli[1,2]
[1]*Dept. of Computer Science, Univ. of California, Davis, CA 95616,*
[2]*Lawrence Berkeley Laboratory, Berkeley, CA 94720*
SNCrivelli@lbl.gov

We describe a novel *ab-initio* method to predict the tertiary structure of new folds. This method uses a two-phase approach: first, it thoroughly samples the conformation space using secondary structure predictions; then, it selects the best models using a combination of different model-evaluation functions. The analysis of the results suggests that this method generates good models although imperfect predictions of secondary structure influence its accuracy. However, these models are successfully selected by the combined scoring function in just a few cases.

Phase I of our method is based on BuildBeta, a ProteinShop[1] tool that automatically creates all possible beta-sheet arrangements from either a prediction file containing the sequence of amino acids and secondary structure prediction from servers[2,3] or a coordinate file in PDB format. When reading a prediction file, BuildBeta generates an extended conformation featuring alpha-helices and beta-strands according to those predictions. This extended conformation is folded into all possible beta sheet arrangements using inverse

kinematics to adjust the flexible backbone in the coil regions and to ultimately decide whether an arrangement is feasible or not. For each arrangement of beta strands into beta-sheets, BuildBeta uses the beta strand alignment scores computed from the tables of Zhu and Braun[4] to decide which residues in the beta-strands should be hydrogen-bonded. Once the sheets are formed, BuildBeta places alpha helices at suitable positions parallel to the constructed beta-sheets.

BuildBeta also permits to pre-specify rigid "core" regions made up by one or more sheets; then, it automatically builds around those pre-specified sheets. This is a possible scenario when predicting proteins that are new folds: the homology modeling servers cannot find any structure that is homologous to the structure of the target but they may find protein structures that have some fragments or cores that are homologous. We use ProteinShop to build an initial partially folded structure for every homologous structure found by assigning to the dihedral angles in the core regions the same values as those in the homologous structure while keeping the rest of the dihedral angles at the ideal values depending on whether they are predicted to be alpha helix or beta strand. The correspondence between the superimposed rigid-body portions is determined by the alignment generated by the meta-server[5]. BuildBeta reads the resulting coordinate file that specifies the beginning and end of the core and automatically extends its partial beta sheets by packing the beta strands that are not part of the core.

BuildBeta's combinatorial approach may generate an enormous number of possible configurations. Phase II selects protein-like models from all initial structures generated in Phase I. First, it uses simple structure validation scores to quickly filter out unreasonable models, trimming the initial pool of models to a more reasonable set. We visually inspect this set and eventually manipulate these structures to create new folds that are added to the pool. Then, a final score combining physical energy scores and statistic scores is applied to further reduce the set of models. Finally, we pick five models among the best ranked ones according to the combined score.

Method Description: Phase I

This phase thoroughly samples the conformation space by generating models with all possible arrangements of predicted beta strands into beta sheets. First, we use the BioInfoBank meta-server[5] -- that uses the 3D-jury consensus approach[6] -- to select those targets that are likely to have new folds. Second, we create one or more consensus secondary structure prediction files according to the secondary structure predictions from the servers[2,3]. BuildBeta reads those prediction files and assigns ideal values to the backbone dihedral angles of residues predicted to be alpha helices and beta strands and uses ProteinShop's inverse kinematics algorithm to rotate and translate the backbone of the flexible coil regions to fully automatically construct beta sheets. BuildBeta creates all possible arrangements of beta strands into beta sheets and uses sequence-matching specificity[4] to align the strands to form hydrogen bonds. BuildBeta arranges helices into likely position parallel to beta-sheets to avoid the collision between secondary structure motifs as well as bury hydrophobic residues. Finally, we use Modeller[7] to refine the coils between strands and helices.

Phase II

This phase implements a filtering and selection procedure to trim the huge set of structures generated in Phase I. First, we use simple model validation scores to remove unreasonable models in order to quickly reduce the size of the initial set. These scores include collision score, compactness score and radius of gyration from Crysol[8] as three independent filters. Second, we develop a combination score based on Dfire[9], RAPDF[10], ProSA[11] and Crysol to select a list of about one hundred candidates from the remaining set. Finally, we select five models according to a combination score and human intuition with convenient interactive operations implemented in ProteinShop.

1. 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.
2. McGuffin,L.J., Bryson,K. & Jones,D.T. (2000) The PSIPRED protein structure prediction server. Bioinformatics **16**, 404-405.
3. SamT06. http://compbio.soe.ucsc.edu/SAM_T06/T06-query.html.
4. 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 Science **8**, 326-342.
5. http://meta.bioinfo.pl/submit_wizard.pl.
6. Ginalski,K., Elofsson,A., Fischer,D., & Rychlewski,L. (2003) 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics. 19(8),1015-1018.
7. Fiser,A. and Sali, A. (2000) ModLoop: automated modeling of loops in protein structures. Bioinformatics **19**,2500-2501.
8. Svergun D.I., Barberato C. and Koch M.H.J. J. (1995) *CRYSOL* - a Program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. Appl. Cryst. **28**, 768-773.

9. Zhou H. and Zhou Y. (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Science, **11**,2714-2726.
10. Samudrala R. and Moult J. (1998) An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J. Mol. Biol. **275**,895-916.
11. Wiederstein M. and Sippl M. J.(2007) ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Research **35**, W407-W410.

PS2-server

$(PS)^2$: protein structure prediction server

C.C. Chen[1], J.M. Yang[1,2,3] and J.K. Hwang[1,2,3]
[1] - Institute of Bioinformatics, [2] - Department of Biological Science and Technology, [3] -Core Facility for Structural Bioinformatics, National Chiao Tung University, Hsinchu, 30050 Taiwan
chieh.bi91g@nctu.edu.tw

In the template-based modeling of protein structures, the template selection and generation for the alignment between the target and the template are the two critical steps, since they will significantly affect the accuracy of the final model prediction. Here, we develop a novel substitution matrix that combines both sequence and structure information for the detection of remote homologs. Our alignment algorithm combines two scoring systems: (1) the S2A2-matrix, a 60×60 substitution matrix based on secondary structure propensities of 20 amino acids. We consider 3 types of secondary structure (α-helix, β-sheet and other); (2) the sequence profile matrix, i. e. PSSM generated by PSI-BLAST[1].

In this study, we applied numerous enhancements and modifications to our previous protein structure prediction server $(PS)^2$ thereby improving the reliability and applicability of the method. The main difference in methodology between the present work and our previous study is to use the S2A2-matrix for the template selection and the alignment between the target and the template. Our previous study utilized a consensus strategy, which combines PSI-BLAST and IMPALA, for these two critical steps. We evaluated the accuracies of the S2A2-matrix on the template selection and the alignment by using Lindahl benchmark and ProSup benchmark, respectively. For the template selection, we compared the S2A2-matrix with other methods on Lindahl benchmark, which consists of 976 proteins, for the fold recognition. The average accuracy of S2A2-matrix (64.1%) is significantly better than the accuracies of PSI-BLAST (34.6%) and profile-profile alignment (56.9% for *prof_sim*). At the superfamily level, the S2A2-matrix, PSI-BLAST, and *prof_sim* identified 75.6%, 25.9%, and 61.3% respectively, of homologous pairs that were ranked in the top 5. At the fold level, the S2A2 matrix, PSI-BLAST, and *prof_sim* identified 54.5%, 4.7%, and 39.6%, respectively, of homologous pairs.

For the alignment between the target and the template, the S2A2-matrix was evaluated on the ProSup benchmark which consists of 127 protein pairs with significant structural similarity but with sequence identity of no more than 30%.

The total numbers of correctly aligned residue pairs of the S2A2-matrix and prof_sim are 9470 and 8009 pairs, respectively. The percentage σ0, the average percentage of correctly aligned residues divided by the length of the structural alignment per protein pair, of the S2A2 matrix, PSI-BLAST, and prof_sim are 58.7%, 35.6%, and 43.6%, respectively. Finally, we evaluated the (PS)2 using the S2A2 matrix for template-based modeling of protein structures. We find that the S2A2 matrix is able to significantly improve the performance on the detection of remote homologous templates. Please note that the key difference between the S2A2-matrix and the other recognition methods is that our method using only sequence information when searching the template database.

In summary, these results demonstrated that the S2A2-matrix is very useful for the template selection and the generation for the alignment between the target and the template. We believe that our approach should be useful in structure prediction and modeling, especially, in detecting remote homologous templates.

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. Chen, C.C., Hwang, J.K. and Yang, J.M. (2006) $(PS)^2$: protein structure prediction server. *Nucleic Acids Res.*, **34**, W152-W157.
3. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*. **292** (2), 195-202.

4. Kabsch,W. & Sander,C. (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical feature. *Biopolymers*. **22**, 2577-2637.
5. Sali,A., Blundell,T.L. (1993). Comparative protein modeling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779-815.
6. 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.

QMEAN
QMEANclust
selfQMEAN
QMEANfamily

QMEAN-based scoring functions for model quality assessment of single models and ensembles

P. Benkert[1], F. Cimarosti[2], T. Schwede[1] and S.C.E. Tosatto[2]
[1] – *Swiss Institute of Bioinformatics, Biozentrum University of Basel, Switzerland*
[2] - *Department of Biology, Universita' di Padova, Italy*
pascal.benkert@unibas.ch

We participated in the quality assessment category of CASP8 with four servers. Two servers operate on single models, namely the composite scoring function QMEAN[1,2] and its derivative QMEANfamily. The other two servers, QMEANclust[2] and selfQMEAN, take into account structural density information contained in the ensemble of models.

QMEAN is a composite scoring function consisting of a linear combination of six structural descriptors: The local geometry is analyzed by a torsion angle potential over three consecutive amino acids. A distance-dependent pairwise Cβ potential as well as an all-atom potential with 167 atom types are used to assess long-range interactions. A solvation potential describes the burial status of the residues. Two simple terms describing the agreement of predicted and calculated secondary structure[3] and solvent accessibility[4] are also included.

QMEANfamily additionally takes into account information from evolutionary closely related proteins of the same family. An ensemble of supplementary models is generated for protein sequences sharing at least 40% sequence identity to the target using the starting model as template. The QMEANfamily score is the average QMEAN score of these models covering the protein family.

QMEANclust combines structural density information provided by the ensemble of models with the QMEAN scoring function as a pre-filter. Only a fraction of the best models ranked by QMEAN scores is used in order to derive the structural density information. The consensus score of a given model is its median GDT_TS to all models in the subset. This approach allows us to counteract the inherent limitations of purely consensus based methods which tend to select models from the most dominant structural cluster thereby missing possible outstanding predictions.

We also investigated whether compiling target-specific statistical potentials based on the models submitted for a given target can improve model selection in a similar way as described by Samudrala and co-workers[6]. The selfQMEAN scoring function is based on specialized statistical potentials which have been trained on the ensemble of models for a given target. The counts extracted from each model are weighted according to the model´s QMEANclust score which increases the influence of more reliable models on the calculation of the statistical potentials.

1. Benkert,P., Tosatto,S.C.E. & Schomburg,D. (2008). QMEAN: A comprehensive scoring function for model quality assessment. Proteins **71**, 261-277.
2. Benkert,P., Schwede,T. & Tosatto,S.C.E. (2008). QMEANclust: Estimation of protein model quality by combining a composite scoring function with structural density information. *Manuscript in preparation.*
3. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* ***292***, 195-202.
4. Cheng J., Randall A., Sweredoski M. & Baldi P. (2005) SCRATCH: a Protein Structure and Structural Feature Prediction Server. Nucleic Acids Research,Web Server Issue **33**, 72-76.

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. Wang,K., Fain,B., Levitt,M. & Samudrala,R. (2004) Improved protein structure selection using decoy-dependent discriminatory functions. BMC Struct Biol 4:8.

RAPTOR

RAPTOR: protein structure prediction by multiple techniques

J. Xu, J. Peng and F. Zhao
Toyota Technological Institute at Chicago
j3xu@tti-c.org

Evolved from a pure threading software package, RAPTOR in CASP8 contains the following modules: a threading module (i.e., the original RAPTOR), an alignment-based model quality assessment program, a multiple sequence alignment module and an ab initio folding program. The strategy used by RAPTOR in CASP8 is as follows. First, a target is threaded to all the templates and the quality of each sequence-template alignment is predicted by a model quality prediction program. According to the predicted GDT score, different methods are employed. If there are multiple templates with very good predicted GDT scores, then we build a multiple sequence alignment between the target and some selected templates, based on which a 3D model is built by MODELLER. If all the templates have very low predicted GDT score, then ab initio folding is employed to build a 3D model. Otherwise, a threading-generated 3D model is directly submitted. The new RAPTOR was not ready for the first 1/3 CASP8 targets at all and was under development during the whole CASP8 season. Therefore, this new RAPTOR was not fully benchmarked yet.

Quality assessment. Different from many model quality assessment methods that directly work on a 3D model, our quality prediction program takes as input an alignment and generates a predicted GDT score. Tested on the alignments generated by RAPTOR for the CASP6 and CASP7 targets, the average prediction error of GDT score is approximately 0.04 and the correlation coefficient between the predicted and real is more than 0.9 for all the alignments and approximately 0.8 for low-quality ones. This quality prediction method is built upon an idea described in [1], which uses SVM to predict the number of correctly aligned positions in an alignment. Current implementation uses better features to describe an alignment and employs an SVM variant to predict the GDT score.

Multiple template method. The top two templates are always used and then we enumerate all the possible combinations of the remaining good templates (at most 5 templates are used in total). For a given set of multiple templates, T-coffee and TM-align are used to generate a multiple sequence alignment from the pairwise alignments produced by threading. Then ProQ and DFIRE are used to rank all the generated models and select the best combination of multiple templates. MetaMQAP is also examined and looks like it is highly consistent with ProQ in model ranking.

Ab initio folding method. We have developed a probabilistic graphical model for ab initio folding, which employs Conditional Random Fields (CRFs) and directional statistics to model the protein sequence-structure relationship. Different from the widely-used fragment assembly method and the lattice model method, our graphical model can explore protein conformations in a continuous space according to their probability. The probability of a protein conformation reflects its stability and is estimated from PSI-BLAST sequence profile and PSIPRED-predicted secondary structure. Experimental results indicate that this new method compares favorably with the fragment assembly method (e.g., Rosetta) and the lattice model (i.e., TOUCHSTONE II). Our method performs well on some hard targets such as T0480, T0495_2 (by Rosetta's domain definition), T0496_1, T0496_2, and T0510_3. A preliminary version of this method is available at [2], in which only a first-order CRF model for conformation sampling is described. Current implementation [3] employs a second-order CRF model for sampling and drives conformation optimization by a simple energy function consisting of Sali's DOPE, Baker's KMBhbond and a simplified solvent accessibility potential ESP.

Finally, a new threading program based on a probabilistic graphical model and a boosting method was also developed in late CASP8 and tested on only several targets [4].

1. Xu, J.. (2005) Fold recognition by predicted alignment accuracy. IEEE/ACM Trans. on Comput Biol and Bioinfo. 2005 Apr-Jun;2(2):157-65.

2. Zhao, F., Li, S., Sterner, B. and Xu, J.. (2008) Discriminative learning for protein conformation sampling. Proteins. 2008 Oct;7 3(1):228-40.
3. Zhao, F., Peng, J., DeBartolo, J., Freed, K.F., Sosnick, T.R. and Xu, J. (2008) A probabilistic graphical model for ab initio folding. Submitted.
4. Peng, J. and Xu, J.. (2008) Boosting protein threading accuracy. Submitted.

RBO-Proteus

De Novo Structure Prediction Using Model-Based Search

TJ. Brunette and O. Brock
University of MassachusettsAmherst
oli@cs.umass.edu

The RBO-Proteus server replaces the Monte Carlo-based search in Rosetta de novo[1,2] with our own search protocol, model-based search (MBS)[3]. We are motivated by the fact that conformational space search is currently viewed as one of the most important obstacles towards accurate protein structure prediction. Our search method differs from existing approaches in that it actively guides conformational space exploration towards promising regions based on information from an all-atom energy function. Our notion of guidance during search is distinct from the concepts of diversification and intensification often used in the context of optimization. Whereas diversification and intensification describe *how* to perform search in a given region of conformation space, our method guides search by choosing *where* to search. The selection of favorable conformation space regions should be guided by the most pertinent and accurate information available. In de novo protein structure prediction the most accurate information comes from the all-atom energy function. While such information previously was deemed too computationally expensive, model-based search is able to guide conformation space exploration with this information without incurring a substantial performance penalty.

Model-based search initially computes a number of short Monte Carlo trajectories. The resulting conformational space samples are analyzed based on their energy and spatial proximity and then clustered into meaningful regions of the search space. These regions are meaningful because they contain samples from Monte Carlo trajectories that with high probability would lead to a single local minimum in the energy landscape. Model-based search is now able to assess the quality of all samples in a region based on the all-atom energy potential. Given a number of regions and an estimate of their likelihood to contain the native conformation, model-based search then guides the exploration of conformation space by selecting which of the regions to search further and how much computational resources to expend per region. Regions are then searched with additional short Monte Carlo trajectories and the process continues for a fixed number of times. By eliminating regions from the ongoing exploration that are unlikely to contain the native structure, model-based search is able to increase the sampling density in the most promising regions, thereby actively guiding search based on highly accurate information about the all-atom energy landscape.

In contrast to most Monte Carlo-based search methods, which treat parallel trajectories as independent, model-based search effectively monitors the progress of these parallel trajectories and aborts some of them in order to restart them in more promising regions of conformation space. This selectively increases the sampling density in promising regions of the search space without the computational burden associated with increasing sampling density over the entire search space.

Due to our integration with Rosetta, model-based search inherits the following algorithmic features. Local search for low-energy conformations starts from an extended backbone conformation. The local, Metropolis Monte Carlo-based search progresses in a number of stages. As the search progresses through the different stages, the move set changes, the number of local search steps are varied, and the accuracy of the energy function is increased. The energy function progresses gradually from a coarse-grained low-resolution energy function that considers secondary structure, residue environment, and inter-residue pairing to a full-atom energy function that includes side chains and solvation effects. Additional details about the move sets, and energy functions can be found in the literature[1,2].

Each iteration of model-based search uses the same move set and energy function as the corresponding stage in Rosetta. The first stage of model-based search occurs after an initial 4,000 Monte Carlo fragment insertions have been attempted for each sample. The remaining 32,000 Monte Carlo steps inside Rosetta are divided into the 13 stages of Rosetta's Monte Carlo-based search. For these stages, the parameters of model-based search are adjusted so that each run finishes in approximately 24 hours on 80 processors. For example, proteins with less than 100 residues use 3,000 extended proteins and five all-atom evaluations to evaluate a region. Proteins larger than 150 residues use 500 extended structures and a single all-atom

evaluation per region. For proteins longer than 250 residues, only the non-all-atom energy function is used. The 5 lowest scoring models are submitted.

1. 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
2. Rohl,C.A., Strauss,C.E.M.., Misura,K.M.S.,& Baker,D. (2004). Protein structure prediction using Rosetta. M*ethods in Enzymology*. **383**, 66-93.
3. Brunette T. & Brock O. (2008) Guiding conformational space search with an all-atom energy function. *Proteins: Structure, Function and Bioinformatics*.

Rehtnap

Experiments in Expectation Maximization for Model Building

R.W. Harrison[1]

[1] – *Department of Computer Science Georgia State Unviersity*

rwh@gsu.edu

Rehtnap is an experimental crazy mixed up server built very quickly before the start of CASP-8 in order to try out using an expectation maximization algorithm (EM algorithm) for protein structure prediction. Instead of aligning sequences, it aligns structures so it is in a sense a "backwards" server. EM algorithms have the potential for merging many different sources of information to produce a model that is as consistent with all of the sources as is possible. This initial implementation suffered from two issues common to EM approaches, namely convergence in the presences of dissonant information and under-prediction in the absence of information. Nonetheless, when it converged (for example T0444), it could perform well. It had the interesting property of predicting small, reasonably accurate fragments when it could not find sufficient information to produce an overlapped set of fragments.

The fundamental idea behind rehtnap is based on the observation that homology between proteins consists of runs of similar sequences without insertions or deletions and gap regions where there is no homology. Sometimes homology between proteins is only seen for short regions and it would be good to use this information in building a model.

The algorithm consists of three phases. First, a profile-profile scan is used to find regions of significant sequence homology. Observationally, a significant homology was defined as a run of at least 10 consecutive local maxima with a z-score of 2.5 or at least 15 consecutive local maxima with a z-score of 2. The z-score was normalized against the variation of the two profiles with each other. Second, each of these alignments was used to extract and convert a fragment of a structure into an initial model where only the homologous atoms were retained. The set of fragments were correlated with each other to produce an overlap matrix and the degree of structural correlation between fragments was used to eliminate structural outliers. Then a graph was constructed that spanned the largest subset of this matrix. Finally, the fragments were iteratively superimposed on the average of the superimposed fragments and the model constructed by energy minimization in a manner similar to that used by panther_server.

The fundamental difficulty rehtnap had was due to dissonant information. Even though individual fragments could superimpose with high precision (rmsd < 1Å for more than 6 residues), the presence of conformation change and regions where there were few samples meant that the convergence of the whole ensemble of fragments to a consistent mean structure could be difficult. Future work will examine both improved approaches to monitoring convergence and handling what is a "multi-modal" solution as well as examining using other computational geometric spaces as a way to enhance convergence.

The other major source of difficulty for rehtnap was its inability to find enough continuous or sufficient sets of fragments. The current algorithm selects the deepest graph that can be derived from the fragment overlap matrix, which is not always the best graph. For example the alignment algorithm easily detected the symmetric structure of T0472 and two "half" graphs were generated from the overlap matrix, but it did not find a fragment that allowed the two half models to be merged into a single model. This property leads to consistent under-prediction. Rehtnap will not predict models without information. Future work will examine how to bring more information to the system by improving or replacing the fragment location algorithms as well as using other techniques to generate or determine fragments.

Protein structure prediction incorporating cotranslation

F.P.E. Huard[1], C.M. Deane[2], J.J. Ellis[1] and G.R. Wood[1]
[1] - *Department of Statistics, Macquarie University, NSW 2109, Australia*
[2] – *Department of Statistics, University of Oxford, Oxford OX1 3TG, UK*
gwood@efs.mq.edu.au

Recent *in vitro*[4] experiments and *in silico*[3,6] computational studies have shown that cotranslation affects the folding pathway of some proteins, especially for ancient folds. To our knowledge, the sequential nature of translation has not yet been incorporated into structure prediction algorithms. SAINT (Sequential Algorithm Initiated at the Nitrogen Terminus) is designed to incorporate cotranslational effects into a fragment-insertion-based protein structure prediction algorithm.

In brief, this first version of SAINT is a cotranslational version of a simplified Rosetta. Residues are extruded from a virtual ribosome and progressively folded; this is done many times and a central configuration selected. Details of the process are now presented.

We simulate translation from the ribosome by iteratively elongating the length of protein to be folded starting from the N-terminus. At each iteration, a fragment of s amino acids is added at the C-terminus of the current model. Fragments are added in a fully extended conformation, with $(\varphi, \psi, \omega) = (-150°, 150°, 180°)$ for all residues. All segments have equal length, except for the last one which has a length between one and the segment length for evident reasons. Each time the chain length increases from ks to $(k + 1)s$ amino acids, the conformation simulated is permitted to change; the $(k + 1)$th fold is an evolution of the kth predicted fold. Structural moves are determined by a folding algorithm, and acceptance of moves is decided using a simulated annealing framework. The final fold of the protein is obtained after all residues are added. The Figure shows an example using the 101 residue long 1qc7 domain. The chain is elongated sequentially using 11 fragments of nine residues and two residues for the last extrusion. This domain was chosen for illustration purposes because it shows evidence for cotranslational folding[2].

We chose a *de novo* method for structure prediction at each step of the elongation process. *De novo* protein structure prediction by fragment assembly has proved successful in recent years (see for example results available on the CASP7 website). Fragment assembly, or fragment recombination, consists of two steps: first, the targeted sequence is divided into overlapping windows of consecutive residues and local fragment structures are assigned to each; second, the local structures are assembled to build models and those of low potential energy are retained[1].

In this first version of SAINT we use Rosetta *ab initio*[7] as a protein structure predictor using fragment recombination. The fragments were built with the Rosetta software; all three- and nine-residue fragments for a window are built from non-homologous proteins of known structure (so producing a fragment library). Rosetta uses a torsion space representation in which models are notated using a list of (φ, ψ, ω) torsion angles. A full atom representation of the model is reconstructed for evaluation by the scoring function. A window in the model is selected randomly and its torsion angles are replaced by the ones of another fragment from the library, thus generating a new tertiary structure.

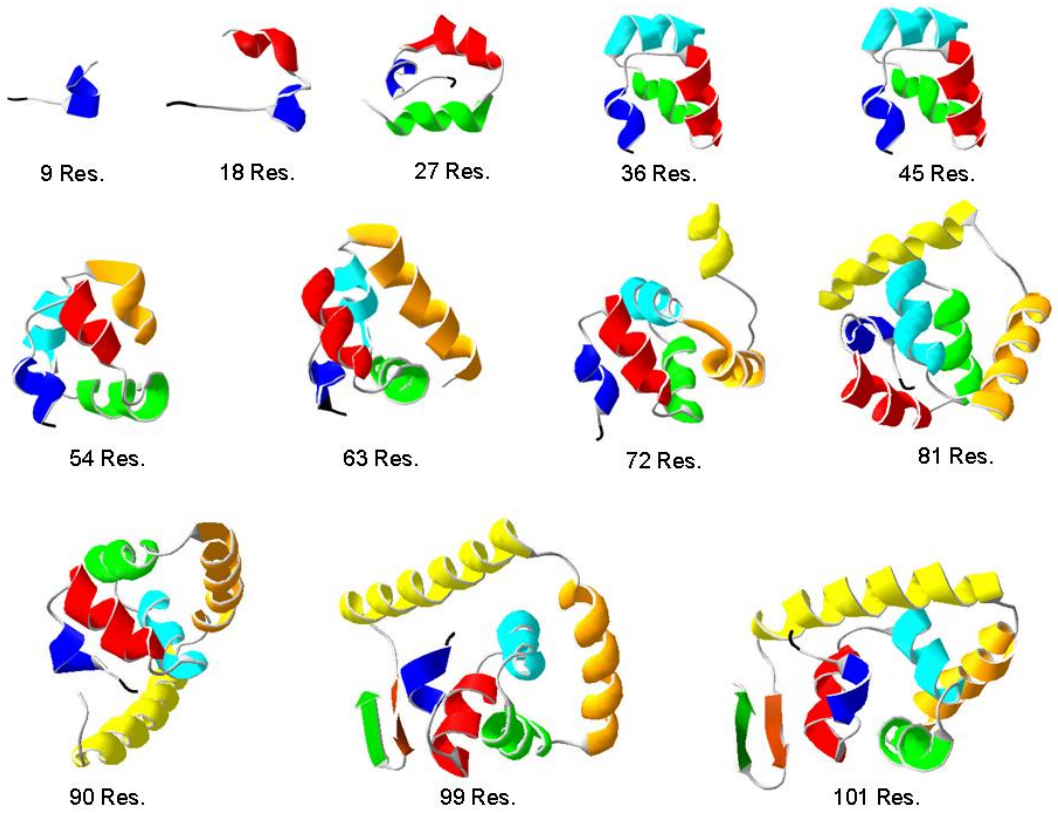


Figure. Cotranslational structure prediction of the 1qc7 domain (101 residues). Residues are extruded in segments of length nine, except for a last segment of two residues. In this example, two hundred runs were carried out; the simulation presented above led to a final structure, among those predicted, which was closest (4.5Å RMSD) to the crystal structure. We can observe that secondary structures form early in the process of elongation, and remain present through to the final extrusion. The N-terminus is represented in black.

The probability of a new tertiary structure, using Bayes' Theorem, is decomposed into a product of two terms: the first reflects the likelihood of the structure regardless of the sequence while the second is a measure of the fitness of the sequence for the given decoy. Final acceptance is determined according to the simulated annealing strategy[7]. We retained the original knowledge based scoring function[5] in which terms describing solvation and electrostatics are based on residue distribution in the structure. Atom steric overlap is penalized in the function, and globally compact structures reflecting favourable Van Der Waals interactions are preferred.

We use a total of 34000 fragment insertion attempts for proteins up to 100 residues in length, and we increase this total on a *pro rata* basis above this figure (e.g. a chain of 143 residues will be simulated by 1.43 times the standard number of insertions). We consider it reasonable to increase the total number of fragment insertion attempts for longer polypeptides. We distribute the total fragment insertion attempts in proportion to the length partially extruded.

For each free-modelling CASP target 1000 predictions were made and the resulting configurations clustered using standard software available in the statistical package R. Selection of a central cluster and then central configurations within this cluster were chosen; the results were examined with standard consistency software and a decision made as to the most likely configuration.

1. Bujnicki, J.M. (2006) Protein-structure prediction by recombination of fragments, *ChemBioChem*. **7**, 19-27.
2. Deane, C.M., Dong, M., Huard, F.P.E., Lance, B.K. & Wood, G.R. (2007) Cotranslational protein folding - fact or fiction? *Bioinformatics* **23**, 142-148.
3. Huard, F.P.E., Deane, C.M. & Wood, G.R. (2006) Modelling sequential protein folding under kinetic control. *Bioinformatics* **22**, e203-e210.
4. Nicola, A.V., Chen, W. & Helenius, A. (1999) Co-translational folding of an alphavirus capsid protein in the cytosol of living cells, *Nature Cell Biol.* **1**, 341-345.
5. Rohl, C.A., Strauss, C.E.M., Misura, K.M.S. & Baker, D. (2004) Protein structure prediction using Rosetta, *Meth. Enzymol.* **383**, 66-93.

6. Sikorski, A. & Skolnick, J. (1990) Dynamic Monte Carlo simulations of globular protein folding, *J. Mol. Biol.* **215**, 183-198.
7. 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.

SAM-T08-2stage

G. Shackelford and K. Karplus
University of California, Santa Cruz

Protein structure prediction continues to be a challenge despite the gains from model builders such as I-TASSER, Rosetta, and undertaker. The best predictions today depend on templates, known protein structures whose sequence is sufficiently similar in part or in whole to the target sequence. These templates provide important constraints in building accurate models. However there are target sequences which have weak or no templates.

To explore possible new approaches to template-free modeling, CASP6 added residue-residue contact predictions as a new category. Subsequently we developed a contact predictor for CASP7 using local structure predictions along with paired statistics including a novel correlation statistic. Its predictions were assessed as the best for CASP7[1]. Since then we have developed a new neural network for CASP8 that employs more inputs[2]. While developing the new predictor, we discovered that by just using local structure predictions, we could build a good predictor. Until then we had assumed that the paired statistics were the main source of predictability and the local structure predictions added only a small amount.

With this new result, we revisited an issue that arises in developing a contact predictor: the sparseness of positive examples. Actual contacts are only about 3% of the total possible pairs of residues. Originally we dealt with the sparseness by reducing the number of negative examples to get a better balance of negative and positive examples while training. The new two-stage predictor resolves this issue by providing a second stage neural network with an enriched set of predictions where the positive examples comprise about 10% of the total examples; no balancing is required.

The first stage uses only local structure predictions and regularized amino acid composition as inputs. We limit resulting predictions to 10*sequence_length. Then paired statistics are calculated for this restricted set of pairs. These statistics along with the log(rank) of the first stage predictions and matching local structure predictions provide the inputs for training a second neural network. The results is a gain of about 3% in overall accuracy.

We also used a new measure, weighted accuracy, that takes the background probability based on separation into account when assessing predictions. In general, the probability that two residues are in contact goes down as the separation increases. The CASP7 assessors dealt with this issue by dividing predictions into three categories: those with separation of 6 or greater, 12 or greater, and 24 or greater.

Using this measure we find that the two-stage predictor may provide better accuracy but lower weighted accuracy. This can be explained if we assume the two-stage predictor making more correct predictions but the predictions have smaller separations than those of a single-stage predictor.

Initially this was a disappointment since we have assumed that accurate contact predictions with large separation would be more useful in model building. However other research [3,4] suggests that there is a need for solutions and constraints involving the super-secondary structure, such as beta sheets and bundled helices. Accurate contact predictions with smaller separation may be more useful than previously anticipated.

1. Shackelford, G. and Karplus, K.. (2007). Contact Prediction using Mutual Information and Neural Nets. Proteins: Structure, Function, and Bioinformatics, 69(S8):159-164.
2. Karplus, K. CASP8 abstract for SAM-T08 predictions.
3. Kuhn, M., Meiler, J. and Baker, D. (2004). Strand-loop-strand motifs: Prediction of hairpins and diverging turns in proteins Proteins: Structure, Function, and Bioinformatics, 54(2):282 – 288.
4. Kaur, H. and Raghava, G.P.S. (2004). Prediction of alpha-turns in proteins using PSI-BLAST profiles and secondary structure information. Proteins: Structure, Function, and Bioinformatics 55 (1):83 - 90

SAM-T08-human

K. Karplus
University of California, Santa Cruz

The SAM-T08 hand predictions used methods similar to SAM_T06 in CASP7 and SAM_T04 in CASP6 [2].

We start with a fully automated method that was essentially the same as the SAM-T08-server, though we froze the code for the server but had several bug fixes and minor improvements for the version used in hand prediction during the summer. The automated method includes improved neural networks for local structure prediction [3] and improved residue-residue contact prediction (see SAM-T08-2stage) [5].

One major change for the method this time was the use of C-beta distance constraints derived from the alignments to templates. These were used to select among the initial alignments and during at least the first run of optimization. The addition of these constraints kept all-alpha structures correctly copied from alignments from being pulled apart by the optimization---a problem that we had in CASP7 and earlier experiments.

After the automatic prediction was done, we examined it visually and tried to fix any flaws that we saw. This generally involved rerunning undertaker with new cost functions, increasing the weights for features we wanted to see and decreasing the weights where we thought the optimization has gone overboard. Sometimes we added new templates or removed ones that we thought were misleading the optimization process. We often did "polishing" runs, where all the current models were read in and optimization with undertaker's genetic algorithm was done with high crossover. These did not usually make much difference to the appearance of the model, but often resolved small clashes or breaks.

Some improvements in undertaker since CASP7 include better communication with SCWRL for initial model building form alignments (now using the standard protocol that identical residues have fixed rotamers, rather than being re-optimized by SCWRL), more cost functions based on the neural net predictions, multiple constraint sets (for easier weighting of the importance of different constraints), and some new conformation-change operators (Backrub and BigBackrub).

We also created model-quality-assessment methods for CASP8, which we applied to the server predictions to get metaserver results. For each target, we did two undertaker optimizations from the top 10 models with two of the MQA methods (SAM-T08-MQAU and SAM-T08-MQAC [1,4]), and considered these models as possible alternatives to our natively-generated models. For some of the targets, we did even more meta-server runs, optimizing from some or all of the server models with various cost functions.

For some targets, we tried breaking the protein up into domains, in an attempt to get more structure searching for domains with few homologs, avoiding contamination of the multiple alignments by neighboring domains.

All results, intermediate files, and working notes are available on the web at http://www.soe.ucsc.edu/~karplus/casp8/

Note: for almost all the targets this summer "we" means Kevin Karplus---students provided some assistance on only 9 targets:T0387, T0388, T0419, T0437, T0443, T0465, T0476, T0484, and T0500.

1. Archie, J.and Karplus, K. Applying Undertaker Cost Functions to Model Quality Assessment. Proteins: Structure, Function, and Bioinformatics accepted.
2. Karplus, K., Katzman, S., Shackelford, G., Koeva, M., Draper, J., Barnes, B., Soriano, M., and Hughey, R. (2005) SAM-T04: what's new in protein-structure prediction for CASP6. Proteins: Structure, Function, and Bioinformatics, 61(S7):135-142.
3. Katzman, S., Barrett, C., Thiltgen, G., Karchin, R. and Karplus, K. Predict-2nd: a tool for generalized protein local structure prediction. Bioinformatics (advanced access 30 Aug 2008). Supplementary material doi:10.1093/bioinformatics/btn438
4. Paluszewski, M. and Karplus, K. Model Quality Assessment using Distance Constraints from Alignments. Proteins: Structure, Function, and Bioinformatics, in press.
5. Shackelford, G. and Karplus, K.. (2007). Contact Prediction using Mutual Information and Neural Nets. Proteins: Structure, Function, and Bioinformatics, 69(S8):159-164.

SAM-T08-server

K. Karplus
University of California, Santa Cruz

The SAM-T08 server predictions use methods similar to SAM_T06_server in CASP7, but with more multiple-sequence alignments, more local-structure alphabets, better calibration of HMMs, and better handling of initial models from alignments by undertaker. Here is a quick overview of the steps:

Use the SAM-T2K, SAM-T04, and SAM-T06 methods for finding homologs of the target and aligning them.

Make local structure predictions using neural nets and the multiple alignments. These neural nets have been newly trained for CASP8 with an improved training protocol. The neural nets for the 3 different multiple sequence alignments are independently trained, so combining them should offer improved performance.

We currently use 15 local-structure alphabets:

STR2	an extended version of DSSP that splits the beta strands into multiple classes (parallel/antiparallel/mixed, edge/center)
STR4	an attempt at an alphabet like STR2, but not requiring DSSP. This alphabet may be trying to make some irrelevant distinctions as well.
ALPHA	an discretization of the alpha torsion angle: CA(i-i), CA(i), CA(i+1), CA(i+2)
BYS	a discretization of Ramachandran plots, due to Bystroff
PB	de Brevern's protein blocks
N_NOTOR N_NOTOR2 O_NOTOR O_NOTOR2	alphabets based on the torsion angle of backbone hydrogen bonds
N_SEP O_SEP	alphabets based on the separation of donor and acceptor for backbone hydrogen bonds
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.
near-backbone-11	an 11-state discretization of the number of residues (represented by near-backbone points) in a 9.65 Angstrom radius sphere around the sidechain proxy spot for the residue.
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 with the amino-acid track having a weight of 1 and the local structure track having a weight of 0.1 (for backbone alphabets) or 0.3 (for burial alphabets). We use these HMMs to score a template library of about 14000 (t06), 16000 (t04), or 18000 (t2k) templates. The template libraries are expanded weekly, but old template HMMs are not rebuilt. The target HMMs are used to score consensus sequences for the templates, to get a cheap approximation of profile-profile scoring, which does not yet work in the SAM package.

We also used single-track HMMs to score not just the template library, but a non-redundant copy of the entire PDB. This scoring is done with real sequences, not consensus sequences.

All the target HMMs use a new calibration method the provides more accurate E-values than before, and can be used even with local-structure alphabets that used to give us trouble (such as protein blocks).

One-track HMMs built from the template library multiple alignments were used to score the target sequence. We plan to use use multi-track template HMMs in future, but we have not had time to calibrate

such models while keeping the code compatible with the old libraries, so the template libraries currently use old calibrations, with rather optimistic E-values.

All the logs of e-values were combined in a weighted average (with rather arbitrary weights, since we still have not taken the time to optimize them), and the best templates ranked.

Alignments of the target to the top templates were made using several different alignment settings on the SAM alignment software.

Generate fragments (short 9-residue alignments for each position) using SAM's "fragfinder" program and the 3-track HMM which tested best for alignment.

Residue-residue contact predictions are made using mutual information, pairwise contact potentials, joint entropy, and other signals combined by a neural net. Two different neural net methods were used, and the results submitted separately.

CB-CB constraints were extracted from the alignments and a combinatorial optimization done to choose a most-believable subset.

Then the "undertaker" program (named because it originally optimized 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 as a frozen core, though in many cases one had much more influence than the others. The alignment scores were not used by undertaker, but were used only to pick the set of alignments and fragments that undertaker would see.

The cost functions used by undertaker rely heavily on the alignment constraints, on helix and strand constraints generated from the secondary-structure predictions, and on the neural-net predictions of local properties that undertaker can measure. The residue-residue contact predictions are also given to undertaker, but have less weight. There are also a number of built-in cost functions (breaks, clashes, burial, ...) that are included in the cost function.

The automatic script runs the undertaker-optimized model through gromacs (to fix small clashes and breaks) and repacks the sidechains using Rosetta, but these post-undertaker optimizations are not included in the server predictions. They can be used in subsequent re-optimization.

1. Katzman, S., Barrett, C., Thiltgen, G., Karchin, R. and Karplus, K. Predict-2nd: a tool for generalized protein local structure prediction. Bioinformatics (advanced access 30 Aug 2008). Supplementary material doi:10.1093/bioinformatics/btn438
2. Shackelford, G. and Karplus, K.. (2007). Contact Prediction using Mutual Information and Neural Nets. Proteins: Structure, Function, and Bioinformatics, 69(S8):159-164.
3. Karplus, K., Katzman, S., Shackelford, G., Koeva, M., Draper, J., Barnes, B., Soriano, M., and Hughey, R. (2005) SAM-T04: what's new in protein-structure prediction for CASP6. Proteins: Structure, Function, and Bioinformatics, 61(S7):135-142.

SAM-T08-MQAO

Model Quality Assessment using Distance Constraints from Alignments

M. Paluszewski[1] and K. Karplus[2]

1 - University of Copenhagen, The Bioinformatics Centre,
2 - University of California, Santa Cruz, Biomolecular Engineering
palu@binf.ku.dk

We present a new approach for addressing the MQA problem. This approach is described in details in the corresponding paper[1]. It is based on distance constraints extracted from alignments to templates of known structure, and is implemented in the Undertaker[2] program for protein structure prediction. Novel features are that we extract non-contact constraints as well as contact constraints and we select the good distance constraints using contact number probability distributions.

The most successful MQA methods in the past have been either consensus methods (looking for features shared by many models in the set) or similarity to a single predicted model[3]. The best MQA algorithm at

CASP7 was Pcons[4] which used a consensus approach where consensus features are extracted from other predictions and used to score the models. The Pcons method therefore needs the predictions from other methods and can not be used to assess the quality of a single model. Our method differs from Pcons since it does not depend on other predictions when the distance constraints are derived from templates.

We use the following steps to extract the distance constraints:

1) Templates and alignments are found using SAM_T06 which is a profile HMM that excels in detecting remote homologs.
2) The distances between pairs of residues in contact are extracted for each alignment found in step 1.
3) For each pair of residues that are in contact in at least one alignment, a consensus distance is computed. After this step, the templates and alignments are therefore reduced to a table of so-called *desired distances* between residues.
4) Weighted constraints are constructed from the desired distances (based on E-values of the templates). Our distance constraints are cost functions that only depend on the distance between two C-beta-atoms and have the lowest cost at the *desired distance* computed in step 3.
5) Our optimization algorithm selects a subset of the distance constraints computed in step 4 using predicted contact number distributions. We select constraints so that residues predicted to have more contacts have more constraints also.
6) Each model is scored according to the distance constraints selected in step 5.

The results shown in in the paper by Paluszewski and Karplus[1] indicate that our method is comparable to the best ranked methods at CASP7 (Pcons and Lee) without using consensus-based methods. When the distance constraints are combined with the other Undertaker cost functions our MQA method can be improved even further as described by Archie and Karplus[5].

The assessments by the SAM-T08-MQAO team is purely done using the distance constraints from alignments described here. Additionally, the distance constraints from alignments have also been applied by the SAM-T08-MQAU and SAM-T08-MQAC teams. Here, the distance constraints are combined with other Undertaker cost functions using the optimization technique described by Archie and Karplus[5]. The distance constraints from alignments are also used for assisting with selection of models and optimization in SAM-T08-server and SAM-T08-human TS predictions.

1. Paluszewski, M. and Karplus, K. (2008). Model quality assessment using distance constraints from alignments. Proteins: Structure, Function and Bioinformatics, (to appear).
2. 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.
3. Cozzetto, D., Kryshtafovych, A., Ceriani, M. and Tramontano, A. (2007). Assessment of predictions in the model quality assessment category. Proteins: Structure, Function, and Bioinformatics. 69(S8):175--183, 6.
4. Wallner, B. and Elofsson, A. (2007). Prediction of global and local model quality in CASP7 using Pcons and ProQ. Proteins: Structure, Function, and Bioinformatics, 69(S8):184―193.
5. Archie, J. and Karplus, K. (2008). Applying Undertaker cost functions to model quality assessment. Proteins: Structure, Function, and Bioinformatics, (to appear).

SAM-T08-MQAU
SAM-T08-MQAC

Using Undertaker's Cost Functions for Quality Assessment

J.G. Archie and K. Karplus
Biomolecular Engineering, UCSC
karplus@soe.ucsc.edu

For structure prediction, Undertaker[1] uses SAM-generated fragments and alignments to generate protein conformations. Conformations most likely to be similar to the real structure are chosen using a combined cost function—a weighted sum of individual cost functions. Each individual cost function measures a characteristic that good predictions should have. For example, some of the most important cost functions for quality assessment measure how well models satisfy distance constraints predicted from alignments[2] and how well models incorporate local structure features, including residue burial[3] and alpha torsion

angles,[4] predicted by neural networks.[5] To address the model quality assessment problem in CASP8, we needed to assign weights to each cost function appropriate for quality assessment. Our approach was to define a measure to judge quality assessment methods, to optimize weights on the Undertaker cost functions to maximize this measure, and to convert Undertaker's combined cost function values to a prediction of GDT_TS in the range [0, 1]. Our measure to judge quality assessment methods and our optimization strategy is summarized here, but is described in detail elsewhere.[6]

The SAM-T08-MQAO group (described in another abstract) used only contact and noncontact predictions from alignments. In addition to the alignment-based constraints, SAM-T08-MQAU included the rest of Undertaker's cost functions and did not include any consensus-based methods. The SAM-T08-MQAC included all Undertaker cost functions as well as some additional consensus terms such as the median TM-score.[7] Median TM-score was calculated by computing the TM-score between a given model and the first model submitted by each server group. The median score is a powerful consensus term.[8] Other consensus terms, calculated in a similar fashion, were median RMSD, median GDT_TS, and median MaxSub.

For a given CASP target, each model has a predicted quality score and an actual GDT_TS score. Two obvious approaches to judge the effectiveness of a quality assessment method are to measure the GDT_TS of the predicted-best model and to measure the correlation between predicted quality and GDT_TS. We adopt an approach that is a hybrid of both of these methods by using a weighted version of Kendall's tau, placing more weight on observations with a better predicted quality. The behavior of the weighted version of Kendall's tau can be altered with a parameter. If the parameter is zero, the measure is equivalent to Kendall's tau; as the parameter approaches infinity, the measure becomes linearly related to the proportion of models with lower quality than the predicted-best model. We subjectively chose a weighting parameter of 3, which places about half of the weight on the predicted-best quarter of models. Our weighted tau is a special case of a weighting described elsewhere[9] that can still be computed in $O(n \log n)$.[10]

For optimization, we chose to maximize an objective function of the average weighted tau for CASP7 targets with solved structures in the PDB. Both human and server models were included in our training set. We used a greedy algorithm devised to select only those cost functions useful for quality assessment. This algorithm started with an empty pool of useful cost functions, and only cost functions in the pool were used to compute predicted model quality. During each iteration, a cost function was chosen which, when added to the pool, increased the value of the objective function by the largest amount. The algorithm stopped when adding another cost function to the pool would fail to increase average correlation by a meaningful amount. For each method, we trained two sets of weights—one for easy targets and one for hard targets—based on the theory that different cost functions might be useful for each category. The difficulty of a target was judged by the e-value of the best SAM alignment.

Finally, on CASP7 data, Undertaker's combined cost function appears to have an approximately linear relationship with GDT_TS. However, to convert the combined cost function to the range [0, 1] we used a standard sigmoidal function with constants set by fitting CASP7 data.

Our results using five-fold cross validation on CASP7 data indicate that our method does quite well in comparison to other methods used in CASP7.[6] We look forward to the CASP8 results to see how our method performs on a new data set alongside state-of-the-art methods.

1. Karplus,K., Karchin,R., Draper,J., Casper,J., Mandel-Gutfreund,Y., Diekans,M. & Hughey,R. (2003). Combining local-structure, fold-recognition, and new-fold methods for protein structure prediction. *Proteins*. **53**, 491-496.
2. Paluszewski,M. & Karplus,K. (2008) Model quality assessment using distance constraints from alignments. *Proteins*. Accepted.
3. Karchin,R., Cline,M. & Karplus,K. (2004) Evaluation of local structure alphabets based on residue burial. *Proteins*. **55**, 508-518.
4. Karchin,R., Cline,M., Mandel-Gutfreund,Y. & Karplus,K. (2003) Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. *Proteins*. **51**, 504-514.
5. Katzman,S., Barrett,C., Thiltgen,G., Karchin,R. & Karplus, K. (2008) Predict-2nd: a tool for generalized local structure prediction. *Bioinformatics*. Accepted.
6. Archie,J. & Karplus,K. (2008) Applying Undertaker cost functions to model quality assessment. *Proteins*. Accepted.
7. Zhang,Y. & Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins*. **57**, 702-710.
8. Qiu,J., Sheffler,W., Baker,D. & Noble,W.S. (2008) Ranking predicted protein structures with support vector regression. *Proteins*. **71**, 1175-1182.
9. Shieh,G.S. (1998) A weighted Kendall's tau statistic. *Stat. Probab. Lett.* **39**, 17-24.

10. Knight,W.R. (1966) A computer method for calculating Kendall's tau with ungrouped data. *J. Am. Stat. Assoc.* **61**, 436-439.

SAMUDRALA

Automated model refinement using knowledge and consensus based restrained torsion angle dynamics

J.A. Horst, C. Mader, R. Samudrala
Computational Biology Group, University of Washington
horst @compbio.washington.edu

Initial models
We select initial starting models from the CASP8 server model set. First we filter out half of the server models by sequentially applying our knowledge based residue specific all atom probability discriminatory function (RAPDF),[1] a van der Waals energy term, the hydrophobic compactness factor,[2] and an electrostatics term, all available within the RAMP suite.[3] With the remaining structures (usually ~120), we apply an iterative density calculation which cycles between a cluster density calculation and removal of outliers.[4] Centroids for the five largest clusters are then selected for further analysis.

RAPDF and consensus based constraint selection
Restrained torsion angle dynamics has been used in many studies to produce highly accurate models from experimentally derived interatomic distance constraints. To exploit this method in structure prediction, a sufficiently accurate and abundant set of constraints is required. We report a method that obtains a distance constraint set sufficient to increase the accuracy of protein structure prediction. Specifically, we use consensus interatomic distances amongst multiple initial predictions, RAPDF to weight the distances, and a compilation method that further improves the set. In regard to RAPDF, we offer a novel and perhaps even more appropriate use for this tool: to select and weight interatomic distances for use as constraints in model refinement. Here our philosophy is that the probabilities derived from a Bayesian analysis of distances observed in a structurally non-redundant database of experimentally derived protein conformations versus random, are likely to be useful to build models similar to the native state and that observed by experiment.

We select interatomic distances for which at least four of the five models show consensus in 0.5 Ångstrom distance bins. We score these consensus distances with RAPDF. Next we apply a batch-by-batch method starting with the highest RAPDF scored consensus distances, to select a single distance for each residue pair. We weight each distance by the RAPDF score and whether they were observed in four or all of the five input models, and finally create three constraint sets using different distance cutoffs (12 Å, 16 Å, 20 Å).

Model building and final selection
The three constraint sets are each used in fifty rounds of CYANA[6] restrained torsion angle dynamics simulations, where a Ramachandran plot is used to approximate probabilities for torsion angles. Each round produces twenty all atom models; in sum three thousand conformations are made, for which we apply the same selection protocol as reported above in "Initial models." In this case, the discriminatory function filter retains one thousand models, which are minimized by ENCAD[7] and side chains are optimized by SCWRL3.0.[8] Finally the iterative density method selects five models for submission to CASP.

Preliminary analysis
Tertiary structure category: The similarity of our final predictive models to conformations produced using experimental data depends strongly upon that of the five input structures. We find that our models are always better than four of the five starting models, observe a mean 0.12 Å CαRMSD improvement, and observe at least eleven cases where the refinements produced better models than any server. The most exciting outcome from this experiment is that the success of this refinement protocol does not depend upon the method used to produce the initial models – it does not matter what method(s) make the input set, our method makes them better. Further, while the final result depends entirely upon the quality of the input models, improvement is seen for initial models of high and low quality. Thus this method is useful for all CASP8 methods for real modeling work.

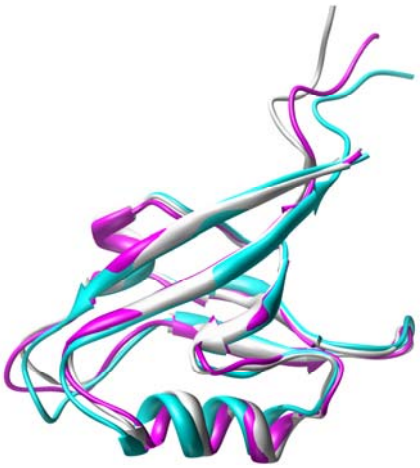
Refinement category: Based on the strengths of NMR to produce the core accurately, and Xray diffraction to produce all atom orientations accurately, we used different structural similarity measures in the iterative density cluster calculations to select our final models: for NMR targets we used TMscore,[9] and for Xray targets we used CαRMSD. For the nine refinement targets which we submitted properly (see "Erratum" below), we observed a dichotomy of measured similarity by TMscore or RMSD, correlating with the

experimental method used to produce the models. Simply, for NMR refinement targets we produced improvement by TMscore but not CαRMSD, and for Xray diffraction targets we produced a mean 0.09 Å CαRMSD improvement but less significant improvements by TMscore. For NMR refinement target TR464 we were excited to see 0.075 TMscore improvement, while this same model is measured by CαRMSD as 4.78 Å further from the experimental conformation than the refinement template. This underscores the importance of similarity measurement in initial and final model selection, which has been discussed by Zhang and Skolnick.[9]

The refinement templates were of a very high quality, which appropriately places a considerable challenge to refinement methods. In accordance with this challenge, we produced 0.36 Å CαRMSD refinement from one of the most accurate templates, TR488, which was already 1.43 Å CαRMSD from the experimental model (see figure).

Automated but not a "server" submission
The torsion angle dynamics simulations (CYANA) are our bottleneck; while we automated the entire process, the time required to sufficiently sample the conformational space prevented us from submitting this as an automated server. Nonetheless, we are in the process of building a server for public use.

Erratum
Around halfway through the experiment we noticed that our filtering method was working in reverse: the best models were being filtered out. Thereafter, depending upon the mean RAPDF score (normalized by length), we either applied the filtering and iterative density method, the iterative density method without filtering, or simply used the five I-TASSER[10] models. Additionally, errors occurred in submitting four of the twelve refinement targets, such that: only one model was submitted for TR453; five errant models were submitted for TR429; and no models were submitted for TR389 or TR454.



Refinement target TR488. Our submitted model (purple) represents a refinement of 0.36 Å CαRMSD from the experimental structure (white), with respect to the 1.43 Å CαRMSD template (cyan)

1. Samudrala R & Moult J. (1997) An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J. Mol. Biol. 275, 893-914.
2. Samudrala R, Xia Y, Levitt M, Huang ES. (1999) A combined approach for ab initio construction of low resolution protein tertiary structures from sequence. In Altman R, Dunker K, Hunter L, Klein T, Lauderdale K, eds. Proceedings of the Pacific Symposium on Biocomputing, 505-516.
3. http://software.compbio.washington.edu/ramp/ramp.html
4. Hung L-H, Ngan S-C, Samudrala R. (2007) De novo protein structure prediction. In Xu Y, Xu D, Liang J, editors. Computational Methods for Protein Structure Prediction and Modeling 2: 43-64.
5. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. (2000) The Protein Data Bank. Nucleic Acids Research. 28, 235-242.
6. Guntert P. (2004) Automated NMR structure calculation with CYANA. Methods Mol Biol. 278, 353-378.
7. 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.
8. Canutescu AA, Shelenkov AA, Dunbrack RL. (2003) A graph theory algorithm for protein side-chain prediction. Protein Science. 12, 2001-2014.
9. Zhang Y, Skolnick J. (2004) Scoring function for automated assessment of protein structure template quality. Proteins. 57, 702-710.
10. Zhang Y. (2008) I-TASSER server for protein 3D structure prediction. BMC Bioinformatics. 9, 40.

Functional site prediction with Meta-Functional Signatures and Homologous ligand-bound structures

J. Horst*, B. Bernard*, S. Iyer, and R. Samudrala
* authors contributed equally to this work
University of Washington
ram@compbio.washington.edu

For a given protein, the residues and their degree of functional importance can be thought of as a signature representing the function. We have recently developed a combination of knowledge and biophysics-based functions to elucidate the relationships between the structural and the functional importance of individual residues and positions[1]. The calculation of such a meta-functional signature (MFS), which is a collection of continuous values representing the functional significance of each residue in a protein, was applied here to the blind prediction of functional sites in the CASP8 experiment.

Meta-functional signature calculation
The protein meta-functional signatures using both sequence and structural information ($MFS^{complete}$) were calculated with predicted structures using sequence & evolutionary conservation, structural stability, and amino acid type scores. Briefly, *sequence conservation score* was calculated from positional relative entropy using amino acid frequencies estimated by a hidden Markov model; *evolutionary conservation score* was calculated by a state to step ratio of residue type changes in a phylogenetic tree built for each position; *structural stability score* was calculated with a residue-specific all atom probability discriminatory function (RAPDF) score as each amino acid was mutated to one of 19 naturally occurring alternatives; and *amino acid type score* was derived from the prior probability of an amino acid being identified as functionally important in two databases of catalytic and ligand binding residues.

Additionally, sequence-only meta-functional signatures ($MFS^{sequence}$) were calculated from the sequence & evolutionary conservation and amino acid type scores to evaluate the ability of the novel method to identify functional sites in the absence of experimental and predicted structures. The $MFS^{complete}$ and $MFS^{sequence}$ scores were calculated for proteins with known ligands noncovalently bound as indicated in the CASP8 experiment (targets T0391, T0395, T0396, T0430, and T0431), as well as all other ligand bound targets in the experiment. To further utilize our predicted structures in the identification of functional sites, we performed a simple spatial clustering of high $MFS^{complete}$ scoring residues.

Inclusion of homologous structures
The spatial clustering of high $MFS^{complete}$ scores occasionally indicated more than one functionally important site. While this may be relevant in a biological context, multiple functional sites are less frequently represented in individual protein structures solved by diffraction or NMR. Therefore, in cases where ligand-bound homologous experimental structures were available and the CASP8 target ligand was identified (with the exception of T0430), this information was used to aid in functional site residue prediction ($MFS^{complete}$ + Homology). To incorporate homology information in functional site prediction, a PSI-BLAST search[2,3] was conducted using the target sequence to identify experimental protein structures with bound ligands. The identified structures were aligned to predicted models with the matchmaker function of UCSF Chimera[4], and the models were refined by energy minimization with the ligand using GROMACS molecular mechanics software[5] and PRODRG ligand parameters[6].

Preliminary analysis
Preliminary analysis has been conducted for proteins with known ligands as indicated in CASP8, and additionally all other ligand-bound CASP8 targets, for which the experimental structure has been released. This is summarized in the table below, with the exclusion of T0395 and T0396 from the known ligand bound set as the experimental structures were not available for analysis.

Method	Targets	Sensitivity	Precision
$MFS^{complete}$	all ligand bound	0.34	0.35
	metal ion	0.78	0.47
$MFS^{sequence}$	all ligand bound	0.35	0.40
	metal ion	0.84	0.43
$MFS^{complete}$ + Homology	known ligand bound	0.59	0.93

The sensitivity is higher for the CASP8 targets with known ligands ($MFS^{complete}$ + Homology) compared to the remaining blind functional site predictions since the ligand geometry assists in identifying contacting residues, whereas $MFS^{complete}$ and $MFS^{sequence}$ identify the most functionally important residues regardless of the specific contacts between a given ligand and protein. The exception is targets with metal ions bound, for which the sensitivity is greatly increased. Fewer residues bind to metal ions, thereby reducing the number of potential false negatives. Additionally, the bonds necessary for coordinating metal ions in functional sites generally arise from the same types of residues involved in catalytic functionality (sought by MFS), such as histidine, cysteine, aspartic and glutamic acid. The only binding residues not identified by MFS for the metal ion targets are those with atoms barely within the 5Å distance range 'binding' definition but not involved in covalent bonding.

Using homology information in binding site prediction gives over two-fold improvement to precision since the number of false positives is reduced as the functional site residues predicted by $MFS^{complete}$ or $MFS^{sequence}$ alone are limited to the known binding site of the homologous structures rather than including clusters in alternative regions of the protein.

Summary and Conclusions

The use of predicted structures to assist automated functional site identification has thus far been an unachieved goal in computational biology. We previously showed that experimental diffraction structures can be exploited in functional site prediction by modeling energetic frustration with substrate absent[1,7]. Here we find that structural minimization procedures remove this signal, such that the performance of $MFS^{sequence}$ is generally better than $MFS^{complete}$. The precision for selection of metal ion binding residues by $MFS^{complete}$ was slightly improved over $MFS^{sequence}$, presumably due to the tight spatial clustering of these sites. This explanation is supported by a correlation between the precision of functional site prediction for ligand bound proteins and the quality of the predicted structures.

The use of structural information present in our predicted models by homologous protein-ligand alignment clearly enhances sensitivity and precision for the targets addressed. We further note that the identity of the bound ligand can make up for global differences in sequence, as the local similarity in geometric and chemical complementarity of proteins is the essence of ligand binding.

1. Wang,K., Horst,J.A., Cheng,G., Nickle,D.C. & Samudrala,R. (2008). Protein Meta-Functional Signatures from Combining Sequence, Structure, Evolution, and Amino Acid Property Information. *PLoS Comput Biol.* **4(9)**, e1000181.
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. Altschul,S.F., Wootton,J.C., Gertz,E.M., Agarwala,R., Morgulis,A., Schäffer,A.A., & Yu,Y-K. (2005). Protein database searches using compositionally adjusted substitution matrices. *FEBS J.* **272**, 5101-5109.
4. Pettersen, E. F., Goddard, T. D., Huang,C.C, Couch,G.S., Greenblatt,D.M., Meng,E.C. & Ferrin,T.E. (2004). UCSF Chimera - a visualization system for exploratory research and analysis. *J Comput Chem*. **25(13)**, 1605-1612.
5. Hess,B., Kutzner,C., van der Spoel,D. & Lindahl,E. (2008). GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J Chem Theory Comput*. **4**, 435-447.
6. Schuettelkopf,A.W & van Aalten,D.M.F. (2004). PRODRG - a tool for high-throughput crystallography of protein-ligand complexes. *Acta Crystallographica*. **D60**, 1355-1363.
7. Cheng,G., Qian,B., Samudrala,R., Baker,D. (2005). Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family evolution using computational design. *Nucleic Acids Research.* **33**, 5861-5867.

A coarse-grained Langevin molecular dynamics approach to *de novo* protein structure prediction

T.N. Sasaki[1], H. Cetin[1] and M. Sasai[1]
[1] – *Department of Computational Science and Engineering, Graduate School of Engineering, Nagoya University*
sasai@tbp.cse.nagoya-u.ac.jp

Our team focuses on *de novo* protein structure prediction. The strategy is based on the Langevin dynamics simulation of a coarse-grained protein chain, in which each amino-acid residue is expressed as one particle [1, 2]. First, we consulted the 3D-jury [3] and other servers to select the FM targets from all released targets. We prepared the fragment candidates for each 9-residue window of each presumed FM target. From these fragment candidates and also from other known protein structures, short and long range interactions among amino-acid residues were constructed to simulate the folding process to model structures: For short-range interactions, we constructed the two-body and multi-body potentials to reproduce the structural tendency of 9-residue fragment candidates. These potentials represent the propensity of secondary structures and other local structures. For long range interactions, we constructed the neighboring-number potential and the beta-sheet potential. The neighboring-number potential expresses the hydrophobic interaction and the exclusive repulsion. This potential was constructed from the known protein structures from which the fragment candidates were abstracted. The parallel and anti-parallel associations of a pair of beta-strands were represented by the beta-sheet potential. The strength of the pseudo-hydrogen bonds between residues in beta-sheets were weighted by using the prediction results of the BETApro [4].

Using this coarse-grained model, the Langevin molecular dynamics simulations were carried out for the selected targets starting from a stretched linear configuration with the simulated annealing method. For smaller targets, a few hundred folding simulations were carried out for each target to find low energy structures. For larger ones, we carried out the folding simulations as much as possible.

From these structures obtained from folding simulations we selected the model structures by using the energy criterion and a newly developed scoring function [1].

1. Sasaki T.N., Cetin H., & Sasai M. (2008). A coarse-grained Langevin molecular dynamics approach to de novo protein structure prediction. *Biochem. Bipphys. Res. Commun.* **369**, 500-506.
2. Sasaki T.N., & Sasai M. (2005) A coarse-grained Langevin molecular dynamics approach to protein structure reproduction. *Chem. Phys. Lett.* **402**, 102-106.
3. Ginalski K., Elofsson A., Fischer D., & Rychlewski L. (2003) 3D-Jury: a simple approach to improve protein structure predictions, *Bioinformatics* **22**, 1015-1018.
4. Cheng J., & Baldi P. (2005) Three-stage prediction of protein beta-sheets by neural networks, alignments and graph algorithms, *Bioinformatics* **21**, Suppl 1;i75-84.

Use of multiplexed replica exchange molecular dynamics with the UNRES force field and distributed computing in ab initio protein-structure prediction

U. Kozłowska[1], A. Liwo[1,2], S. Ołdziej[1,3], R. K. Murarka[1], H. Shen[1], Y. He,[4]
C. Czaplewski[1,2] and H. A. Scheraga[1]

[1] – Baker Laboratory of Chemistry, Cornell University, Ithaca, NY 14853-1301, [2] – Faculty of Chemistry, University of Gdańsk, Sobieskiego 18, 80-952 Gdańsk, Poland,

[3] – Intercollegiate Faculty of Biotechnology, University of Gdańsk, Medical University of Gdańsk, Kładki 24, 80-822 Gdańsk, Poland,

[4] – Biomolecular Physics and Modeling Group, Department of Physics, Huazhong University of Science and Technology, Wuhan 430074, China.

has5@cornell.edu

The structures of the target proteins were predicted using our hierarchical approach[1] in which a polypeptide chain is initially treated at a united-residue level using our UNRES force field and the coarse-grained structures thus found are subsequently converted to all-atom structures.[2,3]

In the UNRES model, 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 energy function was optimized by applying our novel hierarchical optimization method targeted at decreasing the energy while increasing the native-likeness of structures of the training proteins.[2]

To search the conformational space in the UNRES model, we used molecular dynamics which was recently introduced to UNRES[3], enhanced with multiplexed replica exchange (abbreviated MREMD).[4] MREMD searches were carried out at the range of temperatures from T=250 K to T=500 K. To speed up the search for larger proteins, information from secondary structure prediction by PSIPRED[5] was used in the generation of the initial structures. Availability of parallel resources enabled us to treat proteins with size up to 300 amino-acid residues. To extract the candidate conformations from the results of MREMD simulations, we used a procedure developed in our recent work.[2] First, by using the Weighted-Histogram Analysis Method (WHAM), we determined the heat-capacity curves and the folding-transition temperature. Then we chose the analysis temperature as a temperature by 10-20 K lower than the folding temperaure. Using the WHAM results, we calculated the probabilities of all conformations at the analysis temperature, clustered the conformations, and calculated the probabilities of the clusters. The conformations closest to the average structures corresponding to the found clusters were considered as candidate models. The clusters were ranked according to decreasing probability.

1. Scheraga H.A., Liwo, A., Ołdziej, S., Czaplewski, C., Pillardy, J., Ripoll, D.R., Vila, J.A., Kaźmierkiewicz, R., Saunders, J.A., Arnautova, Y.A., Jagielska, A., Chinchio, M. & Nanias, M. (2004) The protein folding problem: global optimization of force fields. *Frontiers in Bioscience* **9**, 3296-3323.
2. Liwo, A., Khalili, M., Czaplewski, C., Kalinowski, S., Ołdziej, S., Wachucik, K. & Scheraga, H.A. (2007) Modification and optimization of the united-residue (UNRES) potential energy function for canonical simulations. I. Temperature dependence of the effective energy function and tests of the optimization method with single training proteins. *J. Phys. Chem. B*, **111**, 260-285.
3. Khalili M., Liwo, A. Rakowski, F., Grochowski, P. & Scheraga, H.A. (2005) Molecular dynamics with the united-residue model of polypeptide chains. I. Lagrange equations of motion and tests of numerical stability in the microcanonical mode. *J. Phys. Chem. B*, **109**, 13785-13797.
4. Nanias M., Czaplewski, C. & Scheraga, H.A. (2006) Replica exchange and multicanonical algorithms with the coarse-grained united-residue (UNRES) force field. *J. Chem. Theory and Comput.* **2**, 513-528.
5. McGuffin L.J., Bryson, K. & Jones, D.T. (2000) The PSIPRED protein structure prediction server. *Bioinformatics* **16**, 404-405.

SELECTpro
3Dpro
FOLDpro
ABIpro

Model Selection using SELECTpro and Tertiary Structure Prediction with FOLDpro, 3Dpro, and ABIpro

A. Randall[1,2], M. Sweredoski[1,2] , P. Rigor[1,2] and P. Baldi[1,2]
[1] – *Institute for Genomics and Bioinformatics,* [2] *– School of Information and Computer Science, University of California, Irvine, CA 92697nstitution2*
pfbaldi@ics.uci.edu

SELECTpro is a purely structure-based model scoring method that participated in the quality assessment category of CASP8. Two servers from our group participated in tertiary structure prediction in CASP8: FOLDpro[1] is a template-based method using a machine learning approach to rank templates and builds models from the top ranked templates. 3Dpro uses the ranked list of templates from FOLDpro to generate a large set of models that are scored and ranked by SELECTpro. The human predictor ABIpro uses the top ranked server model from SELECTpro as input and rebuilds the low confidence portions of the model using the SELECTpro energy function and fragment assembly[2] with simulated annealing.

SELECTpro
SELECTpro scores each model independently using both reduced representation and all-atom energy terms. SELECTpro takes a model backbone as input and by optimizing the side-chain positions using the all-atom energy terms. The reduced representation potential includes terms for predicted secondary structure (SSpro), predicted solvent accessibility (ACCpro), predicted contact map (CMAPpro), local structure independent residue pairing statistical and a novel β-strand pairing treatment. The reduced representation potential also includes statistical terms for local structure independent residue pairing[3], and local structure dependent residue pairing[4]. The all-atom energy function includes terms for hydrogen bonding, electrostatics, solvation effects, and van der Waals interactions. For each target all models are scored and then ranked by the SELECTpro score.

FOLDpro
FOLDpro makes prediction in four steps. First, it extracts pairwise similarity features for a query and all templates in the library using alignment tools and structural feature predictors. It also uses PSI-BLAST[5] to search the query against the template database. Second, a support vector machine (SVM) integrates pairwise features to evaluate the structural relevance of the query and the templates (in the same fold or not). It uses relevance scores to rank the templates. SVM ranking may not always put the best templates on the top of the positive template list. For instance a template in the same fold as the query may be ranked before a template in the same family. So the positive templates are re-ranked by the e-values of PSI-BLAST search if available. Third, FOLDpro generates an alignment between the query and each of the top 5 templates respectively. For templates that can be found by PSI-BLAST, PSIBLAST alignments are used. For harder templates, FOLDpro uses a global profile-profile alignment method COACH[6] to generate the alignments between the query and the templates. Fourth, FOLDpro uses Modeller[7] to build 3D structure for the query, based on its alignments with the templates. Multiple significant templates are combined to generate structures.

1. Cheng,J. & Baldi,P. (2006) A Machine Learning Information Retrieval Approach to Protein Fold Recognition. Bioinformatics. 22, 1456-1463.
2. 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.
3. 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.
4. Zhang,Y., Kolinski,A. & Skolnick,J. (2003) TOUCHSTONE II: A New Approach to Ab Initio Protein Structure Prediction. Biophysical Journal, 85, 1145-1164.
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. Edgar,R.C., & Sjölander,K. (2004) COACH: profile-profile alignment of protein families using hidden Markov Models. Bioinformatics, 20, 1309-1318.

7. Sali,A. & Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol., 234, 779-815.

SHORTLE

Protein structure prediction with statistical potentials and genetic algorithms

D. Shortle
The Johns Hopkins University School of Medicine
Baltimore, MD 21205
dshortl1@jhmi.edu

HOMOLOGY MODELING
The principal focus of the group is energy-based refinement employing a genetic algorithm driven by several statistical potentials[1]. Our original plan was to build several homology models with MODELLER and PSIBLAST derived multiple sequence alignments, then use these models as templates for complete rebuilding with small fragments of protein structure selected on the basis of local side-chain and backbone energies[2]. However, this approach proved to be too slow to keep pace with the rate of target release, so instead the tarball of server models provided by the CASP website was downloaded, and all models with CG atoms (approximately 180-240) served as the input structures for refinement.

Almost all targets of length less than 200 amino acids were attempted, regardless of classification as "human/server" or "server only". A single refinement protocol was applied. Briefly, the genetic algorithm program loads the model structures, adds heavy atom side-chains, does a grid search of major rotamers at each position, and a list of energy terms are scored. The N best structures (ranked by a simple sum of z-scores for a specified list of parameters) are selected as the initial population. Two parents are selected at random, a child is formed by a single random cross-over, a rotamer grid search is carried out along with an energy minimization protocol involving changes in phi/psi/omega of turn residues. The child structure is saved if it satisfies a specified improvement in the selected energy parameters. After the population of structures has increased by N children, the 2N → 1N selection restores the base population to N. A list of tactics are employed to prevent premature convergence.

To reduce the rate of random loss of diversity, the genetic algorithm is run as a series of epochs, consisting of 3 to 5 generations, with the best structures after the final generation saved to file. Multiple cycles through the same epoch accumulates an ensemble of structures with different, more-or-less random subsets of backbone structure retained. Epoch1 consisted of 3 generations with N = 100, initialized by 100 randomly selected structures from the tarball, with selection for atomic solvation and local atomic interactions; 40 structures are saved after these 3 generations. Epoch2 consisted of 4 generations, N = 200, initialized with randomly selected structures from epoch1; parameter selection was similar to epoch1 plus total atom-pair energy. Epoch3 consisted of 5 generations with N = 200, with only total atom-pair, hydrogen-bond, solvation, and compactness selected. The submitted model had the best sum of z-scores for these parameters.

NEW FOLD PREDICTION
The PSIPRED server is used to predict secondary structure, and the conserved hydrophobic residues in the PSI-BLAST multiple sequence alignment are assigned "core" positions, whereas residues that are charged or highly polar in more than 40 percent of homologues are assigned "surface" positions.

The sequence is divided into large overlapping fragments (30-60 residues in length) that start and end at the ends of predicted helices/strands. Approximately 2000 structures for each of these super-secondary structural fragments are constructed from pieces of backbone structure 3 to 8 residues in length taken from a collection of x-ray structures based on local statistical potentials[2] without regard to the template's amino acid sequence. These constructed fragments are recombined in all possible orders to generate full length decoys low in energy (side-chain interactions plus solvation) and with favorable structural heuristics (compact hydrophobic core, turn relationships, etc.), which are then clustered and submitted to refinement by genetic algorithm. For some targets the super-secondary structural fragments were submitted to epoch1 of GA prior to assembly into full length decoys. Major mistakes were made in the treatment of beta strand interactions.

1. Fang, Q & Shortle, D (2006) Protein refolding in silico with atom-based statistical potentials and conformational search using a simple genetic algorithm J. Mol. Biol. 359:1456-1467.
2. Fang, Q & Shortle, D. (2005) A consistent set of statistical potentials for quantifying local side- chain and backbone interactions. Proteins 60: 90-96.

SiteHunter

A threading-based approach for the prediction of protein ligand and protein-DNA interactions

M. Brylinski, M. Gao and J. Skolnick
Center for the Study of Systems Biology, School of Biology, Georgia Institute of Technology, 250 14th Street NW, Atlanta, GA 30318
skolnick@gatech.edu

In the post-genome era, the rapid accumulation of protein sequences with unknown structure and function has motivated the development of computational tools for protein structure and function prediction at the proteome scale. Here, we present SiteHunter, an automated webserver for the prediction of protein-ligand and protein-DNA interactions through protein threading. SiteHunter comprises two method components: FINDSITE[1] that detects binding pockets for small molecules, and DBD-Hunter[2] that identifies DNA-binding sites.

Protein threading is capable of detecting remote, yet evolutionary related homologues. The conservation of functional sites among homologous proteins allowed us to develop FINDSITE, a highly accurate method for ligand-binding site prediction and functional annotation. FINDSITE employs template identification, structure superimposition and binding site clustering as follows: First, for a given target sequence, structure templates are selected by three threading procedures: PROSPECTOR_3[3], Sparks2[4] and SP3[5]. Subsequently, template structures bound to ligands are identified and superimposed onto the target protein structure using the structural alignment algorithm TM-align[6]. In CASP8, we used TASSER[7] models as the reference structures. After the superimposition, putative binding sites are inferred through the clustering of the template ligands, and the predicted sites are ranked according to the number of templates that share a common binding pocket. Considering a cutoff distance of 4Å as the hit criterion, benchmarks carried out for weakly homologous TASSER models demonstrated a success rate of 67.3% for identifying the best of top five predicted ligand-binding sites with a ranking accuracy of 75.5%. The median sensitivity and Matthew's correlation coefficient (MCC) between predicted and observed binding residues are 0.64 and 0.59, respectively. FINDSITE tolerates structural inaccuracies in protein models up to a RMSD of 8-10Å from the native structure. Furthermore, by exploiting the chemical properties of template-bound ligands that occupy predicted binding pockets, FINDSITE constructs ligand templates for use in ligand-based virtual screening. Each of the CASP8 targets was screened against the KEGG compound library and the top-ranked molecules were included as a part of the function prediction.

The second component of SiteHunter carries out the DNA-binding function prediction. This component utilizes a knowledge-based method, DBD-Hunter, which requires either the structure or sequence as the input for a given target. In CASP8, we applied the prediction procedure that requires only the sequence input. The method combines structural template threading and a statistical pair potential derived from known DNA-protein complexes. First, the target sequence is threaded against a template library composed of DNA-protein complex structures with the threading program PROSPECTOR_3. For template hits with a Z-score better than a threshold value, the statistical potential energy between the target protein and the template DNA is calculated by evaluating DNA-protein contacts. The templates are then ranked according to their interfacial energies. If the lowest interfacial energy is below an energy threshold, the target is predicted to be a DNA-binding protein. For predicted DNA-binding proteins, we further infer DNA-binding protein residues from their templates. In benchmark tests on 179 known DNA-binding proteins and 4000 non-DNA-binding proteins, this method archives a sensitivity of 62% at a precision of 80% in the function prediction; and the mean sensitivity and MCC for binding residues prediction are 0.67 and 0.59, respectively.

1. Brylinski,M. & Skolnick, J. (2008). A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proc Natl Acad Sci U S A* **105**, 129-34.
2. Gao,M. & Skolnick,J. (2008). DBD-Hunter: a knowledge-based method for the prediction of DNA-protein interactions. *Nucleic Acids Res* **36**, 3978-92.
3. Skolnick,J., Kihara,D. & Zhang,Y. (2004). Development and large scale benchmark testing of the PROSPECTOR_3 threading algorithm. *Proteins* **56**, 502-18.
4. Zhou,H. & Zhou,Y. (2004). Single-body knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. *Proteins* **55**, 1005-13.
5. Zhou,H. & Zhou,Y. (2005). Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. *Proteins* **58**, 321-8.

6. Zhang,Y. & Skolnick,J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* **33**, 2302-9.
7. Zhang,Y., Arakaki,A.K. & Skolnick,J. (2005). TASSER: an automated method for the prediction of protein tertiary structures in CASP6. *Proteins* **61** Suppl 7, 91-8.

SMEG-CCP

Prediction of native contacts, 3D structures and model quality using consensus contacts

B.H. Stehr, M. Lappe
Max-Planck-Institute for Molecular Genetics, Berlin, Germany
{stehr, lappe}@molgen.mpg.de

The SMEG-CCP approach (Sample MEan of Graphs Consensus Contact Prediction) uses contact information derived from server models to predict residue-residue contacts, 3D structures and model quality.

For each target, the server predictions marked as model 1 were converted to contact maps using the CASP contact definition of 8Å between C-beta atoms (C-alpha for glycins). The sample mean[1] of these contact maps contains for each contact the frequency of occurrence in the input ensemble. The frequencies were ranked in descending order and the top *n* were submitted as predicted contacts where *n* is the expected number of contacts in the given target. To determine *n*, we again used a consensus approach choosing *n* as the median number of contacts predicted in the server models.

In many cases, the predicted contact maps are better in terms of prediction accuracy and coverage than the best input model (see e.g. Figure 1).

From the predicted contact maps we calculated 3D models with the DISTGEOM[2] program from the TINKER[3] package. Each contact was translated into a distance restraint with 2.6Å lower and 8Å upper bound between C-beta atoms.

The quality of server models was predicted based on the agreement of their contact map with the sample mean. For each contact present in a model, the number of structures in the ensemble that share that same contact gives an estimation of the likelihood of this contact being native. These values summed over all contacts in a model give the raw quality score for the model.
Raw scores from a training set were fitted to GDT scores to derive quality estimates in terms of GDT.

The method works particularly well for medium-difficulty targets where enough consensus information is available but agreement between models is not too high.

In a real-world setting, the input models for this method can be obtained from structure- or contact prediction servers.

Figure 1: Consensus contact prediction of T0409. The prediction (b) is derived from the ensemble (a) by the method described above. The native contact map (c) is shown for comparison.

1. Jain, B. and Obermayer K. (2008). On the sample mean of graphs. In *IJCNN 2008 Conference Proceedings. 993-1000*

2. Hodsdon,M.E., Ponder, J.W. and Cistola,D.P. (1996). The NMR Solution Structure of Intestinal Fatty Acid-binding Protein Complexed with Palmitate: Application of a Novel Distance Geometry Algorithm. J. Mol. Biol. 264, 585-602.
3. Ponder,J.W. and Richards,F.M. (1987). An Efficient Newton-like Method for Molecular Mechanics Energy Minimization of Large Molecules, J. Comput. Chem. 8, 1016-1024.

Softberry

Softberry tools for protein structure modeling and docking

V. Solovyev [1,2], D. Affonnikov[2], A. Bachinsky[2], N. Bakulina 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 are continuing development of a suite of structure prediction programs that were applied to analyze CASP8 targets. Most of these programs can be used within window based **Molquest *(http://ww.molquest.com)*** computer package or run on the web server at http://**www.softberry.com**. Identification of disordered regions in proteins was made by the ***Pdisorder*** program that uses a combination of neural network (NN), linear discriminant function (LDF) and a smoothing procedure. At the first stage, we compute features in a sliding window of 31 residues for neural network and for the linear discriminant function At the second stage, we apply a smoothing procedure that computes chances for the positions of query sequence to be in ordered regions.

Initial step in 3D modeling is selection of a template structure for a query sequence, or selection of a set of short 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 structure similarity of a query and the template proteins, as well as compatibility of query with solvent accessibility of template protein. Secondary structure of a query protein is predicted by our ***SSPRED*** 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 a few template-query pairs are selected for further modeling.

Building side chain and loop coordinates for a query protein based on the 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). Unfavorable conformations are then filtered out using several single-residue criteria, pairwise VDW interaction energy, and Goldstein DEE algorithm [1]. 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 built by ***Getatoms*** program are further refined by ***MDynSB*** program, which performs energy minimization using AMBER force field (2) and optimization of a protein structure via molecular dynamics or optimization and folding of a protein via simulated annealing protocol in an implicit water solvent. Recently we have developed a useful extension of this program***: MDdoc*** that realize ab initio docking of a molecular ligand or peptide by exhaustive cavity search with global optimization. The developed method of blind docking shows a good accuracy in prediction of the native binding modes of flexible ligands. At the test set of 8 ligands the method achieved 100% accuracy, i.e. the native binding mode are found as the mode with highest binding energy calculated in the all-atom force field.

1. Goldstein R.F. (1994) Efficient rotamer elimination applied to protein side-chains and related spin glasses. Biophys J. 66, 1335-1340.
2. 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.

From comparative modeling to *de novo* folding with Phyre, Poing and Phragment

L.A. Kelley[1], B.R. Jefferys[1] and M.J.E. Sternberg[1]
[1] – *Structural Bioinformatics Group, Department of Life Sciences, Imperial College London,SW7 2AY, United Kingdom*
l.a.kelley@imperial.ac.uk

The Sternberg group entered four servers for structure prediction: *Phyre_de_novo*, *Phyre2*, *Phragment* and *Poing*. All four servers share a common core for homology detection. A weekly-updated HMM fold library based on SCOP and the PDB is constructed using PSI-Blast[1] and a reduced redundancy sequence database (UNIREF50). Incoming query sequences are similarly processed to generate an HMM which is searched against the fold library using HHsearch[2]. Where HHsearch fails to detect confident matches we perform an homology network data mining procedure based on SCOOP[3]. These processes provides a limited list of potential templates and associated confidence scores. When a selected template leaves >30 residues of the query unmodelled, such regions are extracted, placed in the Phyre pipeline and the process repeated until the entire length of the query is modelled. Individually modelled domains are then reconnected using a modified fragment assembly protocol.

The *Phyre2* system simply selects the top five templates by confidence score, performs loop modeling on missing regions using a loop library and cyclic coordinate descent (CCD)[4] and adds sidechains using the R3 algorithm[5] and the sidechain rotamer library of Dunbrack[6]. For the *Phragment* server, if the top scoring template is less than 95% confident and the query is <150 residues in length, an in-house fragment assembly *de novo* prediction is performed.

Poing is a fast new model for template-free protein structure prediction based upon Langevin dynamics with novel models for physicochemical effects. Three features of protein folding are modelled in a novel way. 1) The repulsive steric interaction between two particles depends upon the probability that atoms in an all-atom model of those particles a given distance apart would clash sterically, based upon analysis of sidechain and backbone conformations in the PDB. 2) The polar interactions of the backbone (i.e. hydrogen bonds) are modelled by initially calculating the likely position of the O and H atoms involved in the interactions. Forces between the associated backbone particles aim to bring the notional O and H atoms closer together. 3) We have enhanced the standard implicit solvent model of the Langevin equation by ensuring that drag and kicks only act upon parts of a particle exposed to solvent. This ensures that the internal parts of a protein are not subject to solvent effects, a key advantage of modelling an explicit solvent. In addition, kicks are more frequently initiated against hydrophobic residues which in turn drives hydrophobic collapse. The simplicity of the model leads to very fast folding. Proteins up to 90 residues can fold to a stable state in 5-20 minutes. Our current predictions use 100 fold replicates and require between 8 and 30 CPU hours. As with the *Phragment* server, low confidence templates (<95%) that are less than 350 residues in length (*Poing* is fast enough to process longer sequences than *Phragment*) are sent through the *Poing* folding system.

Finally, the *Phyre_de_novo* server is our method to combine the best aspects of the three other servers. First, if multiple high confidence templates are detected by HHsearch or SCOOP, models are built from each template and structurally clustered. The models from the largest cluster are used to define the templates that will be used in a multiple template modeling phase using Modeller. In the absence of high confidence templates, the *Poing* folding system is used. Finally, large insertions or long unmodelled regions at the N- and C-termini are modeled by *Poing* in the context of the predicted model which is held fixed.

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. Söding J. (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics 21, 951-960.
3. Bateman A & Finn, R.D. (2007) SCOOP: a simple method for identification of novel protein superfamily relationships. Bioinformatics 23(7):809-814
4. Canutescu, A.A. & Dunbrack RL. (2003) Cyclic coordinate descent: A robotics algorithm for protein loop closure. Protein Science, 12:963-972.
5. Xie, W. & Sahinidis, N.V. (2006) Residue-rotamer-reduction algorithm for the protein side-chain conformation problem, Bioinformatics, 22(2), 188-194
6. Dunbrack Jr., R.L. (2002) Rotamer libraries in the 21st century. Curr. Opin. Struct. Biol. 12, 431-440.

7. Sali, A. & Blundell, T.L.. (1993) Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234, 779-815

Sternberg

Using ConFunc for binding site predictions in CASP8

M.N. Wass and M.J. Sternberg
Structural Bioinformatics Group, Imperial College London
mark.wass04, m.sternberg @imperial.ac.uk

ConFunc[1] is a sequence based function prediction algorithm that identifies conserved residues associated with individual Gene Ontology (GO) functions. Our approach during CASP8 was to test if we could use the function predictions made by ConFunc to further infer binding site residues. ConFunc was run as two servers, the first ConFunc1D used solely sequence information to infer functional residues, while ConFunc3D incorporated structural data into the prediction process. The ConFunc approach was also extensively used for the human predictions made by the Sternberg group.

The prediction of binding site residues is an important task, which while related to the prediction of function is also distinct from it. It was therefore necessary to modify the ConFunc approach for the prediction binding site residues. For a query sequence ConFunc identifies homologues using PSI-BLAST[2]. Those with GO annotations are further aligned using MUSCLE[3]. This alignment is split into subalignments with each subalignment representing a different GO function. Conserved positions are identified for each subalignment; Position Specific Scoring Matrices are generated for them and used to calculate an e-value for the similarity between the query sequence and each of the subalignments. The e-values associated with each GO term are finally used to infer function. ConFunc uses the identification of conserved residues to infer function, so in CASP8 we have taken this a step further by using the function prediction to direct the prediction of functional residues.

The standard ConFunc prediction consists of a set of GO terms, so to predict binding sites for CASP, the most functionally specific predicted GO function was used as a starting point. The ConFunc1D server predictions were simply the conserved positions in the subalignment for this function. Residues are conserved for functional roles, for example catalytic residues, however residues are also conserved for structural reasons and may be essential for the protein to maintain its fold. So the approach of ConFunc1D predicting all of the conserved residues as part of a binding site, is likely to include residues that are not present in a binding site and are conserved for other reasons.

The ConFunc3D server attempts to refine the predictions made by the 1D server by incorporating structural information into the prediction process. The aim of this is to distinguish between residues conserved for functional and structural reasons. This is done by considering the solvent accessibility of the conserved residues and secondly their spatial location relative to one another. ConFunc3D ran BLAST to search for homologues of the target in the pdb to provide a template structure. The conserved residues from the subalignment were mapped onto the protein structure. Any conserved residues with a solvent accessibility less than 5Å^2 (calculated using HSSP) were removed from the prediction, with the aim of removing residues that are completely buried in the core of the protein and therefore unlikely to be binding a ligand. A simple clustering of the residues was then performed such that only conserved residues within 5Å of another conserved residue were retained for the prediction. This step removes conserved positions that are isolated on the template structure and as ligands generally bind to multiple residues.

Finally we made manual predictions, which in many cases were based upon the server predictions. We additionally used predicted structures from our Phyre[4] servers and other server structure predictions for mapping of the residues as opposed to the identification of a template in the pdb.

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. 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.
3. Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Res*, **32**, 1792-1797.
4. Bennett-Lovsey, R.M., Herbert A.D., Sternberg M.J., Kelley L.A. (2008) Exploring the extremes of sequence/structure space with ensemble fold recognition in the program Phyre. Proteins **15**, 611-625.

StruPPi

Comparative modeling using consensus information from multiple templates and physics-based refinement

B.H. Stehr[1], J.M. Duarte[1], I. Filippis[1], S. Rajagopal[1], K. Syal[1], S. Risbud[1], L. Holm[2] and M. Lappe[1]

[1] - *Max-Planck-Institute for Molecular Genetics, Berlin, Germany*
[2] - *Institute of Biotechnology, University of Helsinki*
{stehr, lappe}@molgen.mpg.de

We implemented a homology modeling pipeline to address the following questions: For which cases does the use of consensus information from multiple templates improve model quality compared to single template modeling?
Can the predicted models and in particular regions with weak consensus be improved by physics-based refinement procedures?

The pipeline consisted of eight basic steps of which six are fully automatic:

1. Identify potential templates using a BLAST and PSI-BLAST search against a non-redundant PDB subset.

2. In cases where no obvious templates can be found, use the Global Trace Graph[1] method to detect remote homologs.

3. Manually select templates using template clustering, secondary structure prediction[2] and consensus information.

4. Align selected template structures and extract consensus contacts and dihedral angles.

5. Align the target sequence to a profile derived from the template alignment.

6. Build homology models based on consensus contacts and angles using the DISTGEOM[3] program from the TINKER[4] package.

7. Refine models with simulated annealing and molecular dynamics simulations[5].

8. Manually choose one or more models out of different sets of input templates and different refinements schemes for submission.

We tried different all-atom refinement schemes, including simulated annealing and molecular dynamics simulations with different force fields and parameters. Preliminary results suggest that while local structure in regions of low consensus can be improved, overall model quality in terms of GDT-TS only gains minimally on average.

Our method for target to multiple template alignment using a new multi-body potential was not ready in time for CASP8. As a consequence, the alignment quality in step 5 suffered in many cases and yielded suboptimal models.

1. Heger,A., Mallick,S., Wilton,C., Holm,L. (2007). The global trace graph, a novel paradigm for searching protein sequence databases. *Bioinformatics* ***23***(18), 2361-2367.
2. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* ***292***, 195-202.
3. Hodsdon,M.E., Ponder,J.W. and Cistola,D.P. (1996). The NMR solution structure of intestinal fatty acid-binding protein complexed with palmitate: application of a novel distance geometry algorithm. *J. Mol. Biol.* ***264***, 585-602.
4. Ponder,J.W. and Richards,F.M. (1987). An efficient newton-like method for molecular mechanics energy minimization of large molecules, *J. Comput. Chem.* ***8***, 1016-1024.
5. Lindahl, E., Hess, B. and van der Spoel, D. (2001). GROMACS 3.0: A package for molecular simulation and trajectory analysis. J. Mol. Mod. 7, 306-317.

Protein residue contact prediction by SVMSEQ and LOMETS servers

S. Wu and Y. Zhang
Center for Bioinformatics and Department of Molecular Bioscience, University of Kansas, 2030 Becker Dr, Lawrence, KS 66047
yzhang@ku.edu

Pair-wise residue-residue contacts in proteins can be predicted from both threading templates and sequence-based machine learning. Accordingly, protein targets can be categorized into two groups. For proteins with reliable threading templates, the accuracy of contact prediction collected from the templates dominates that from sequence-based *ab initio* prediction. For the targets where threading could not identify meaningful templates, the sequence-based contact methods are anticipated to generate better contact prediction than the template-based ones. Apparently, different methods are needed for treating different protein targets.

We developed two contact prediction methods: LOMETS[1] and SVMSEQ[2]. LOMETS is a meta-threading server with 9 locally-installed threading programs to generate threading alignments, i.e. FUGUE[3], HHSEARCH[4], PAINT, PPA-I, PPA-II[1], PROSPECT2[5], SAM-T02[6], SP3 and SPARKS2[7,8]. Contact predictions in LOMETS are generated by collecting those contact residue pairs which occur more frequently in the top-scoring threading templates.

SVMSEQ is machine-learning-based *ab initio* contact prediction method which trains a variety of sequence-derived features on contact maps[2]. The features used by SVMSEQ include (1) *Local window features:* position-specific scoring matrices, secondary structure predictions, solvent accessibility predictions and (2) *In-between segment features:* the contact order, the compositional percentage of three different secondary structure elements and two different burial states for the in-between residues, state distributions of the in-between residues, and the local features of five selected in-between residues. Overall, for short/medium/long ranges (corresponding to the sequence separation in 6-11, 12-23 and ≥24 residues, respectively), there are 781/787/918 input features, which are used for SVM[9] to classify the contact and non-contact pairs.

In CASP8, we combine the two methods by assigning a consensus score for each contact pair, i.e.

$$\text{Score}=\text{Score}_{\text{svmseq}}+w*\text{Score}_{\text{LOMETS}}, \quad (1)$$

where $\text{Score}_{\text{svmseq}}$ is the confidence score returned by SVM training, $\text{Score}_{\text{LOMETS}}$ is the relative frequency of the contacts appearing in the threading templates, and w is a weighting factor. Based on the training results of 105 proteins, we use w=2.2 for "Easy" and "Medium" targets and w=0.6 for "Hard" targets. The category of Easy/Medium/Hard is decided by LOMETS[1]: if there is at least one template with Z-score>Z_0 in each of the threading programs, the target is Easy; if there is no good template of Z-score>Z_0 in any of the programs, it is Hard; the others are Medium targets[1], where Z_0 is a program-specific Z-score cutoff for generating confident predictions.

In general, the contact predictions with a higher consensus score will have a higher accuracy. For example, in our benchmark test, if a contact prediction with a confidence score >0.32, the average accuracy is >40%. But in CASP8, we submit all our contact predictions (up to 60 times of target length) with the contact pairs ranked by the confidence score in (1).

1. Wu, S. & Zhang, Y. (2007). LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic acids research* **35**, 3375-3382.
2. Wu, S. & Zhang, Y. (2008). A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. *Bioinformatics (Oxford, England)* **24**, 924-931.
3. 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.
4. Soding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics (Oxford, England)* **21**, 951-960.
5. Xu, Y. & Xu, D. (2000). Protein threading using PROSPECT: design and evaluation. *Proteins* **40**, 343-354.
6. 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.
7. 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.

8. Zhou, H. & Zhou, Y. (2005). Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. *Proteins* **58**, 321-328.
9. Joachims, T. (2002) Dissertation: Learning to Classify Text Using Support Vector Machines. Software available at http://svmlight.joachims.org/.

TASSER

TASSER-based protein structure prediction in CASP8

H. Zhou, S. Y. Lee, S. B. Pandit and J. Skolnick
Center for the Study of Systems Biology, School of Biology
Georgia Institute of Technology, 250 14th Street, N.W., Atlanta, GA 30318
skolnick@gatech.edu

The TASSER human-expert structure prediction group implemented the previously developed TASSER methodology[1] along with the tertiary restraints derived from models generated by METATASSER, pro-sp3-TASSER and other CASP8 servers. CASP8 targets are classified as Easy, Medium or Hard when the Z-score of the first SP^3 threading template is > 6.0, $4.5 \leq$ Z-score ≤ 6.0, or < 4.5, respectively. For Easy targets, CASP8 server models were ranked using TASSER-QA [2], and the top 20 models are used to derive the distance and contact restraints for TASSER refinement. Then, the top five cluster centroid structures from SPICKER [3] are selected. For Medium/Hard targets, in addition to the previously used chunk-TASSER approach [4, 5], an upgraded version of TASSER with improved contact predictions (TASSER 2.0 [6]) was applied. We performed multiple simulations of chunk-TASSER and TASSER 2.0, followed by SPICKER clustering. The generated models are then ranked by TASSER-QA, and the top five centroid models are selected. Human intervention was mainly employed for possible multiple domain, Medium/Hard targets. In this case, we empirically obtained the domain boundaries for the target sequence and modeled the domains separately. The domains are then combined to generate full length models using short TASSER simulations. For all targets, the main-chain and side-chain atoms are built from the cluster centroid structures using PULCHRA[7].

1. Zhang, Y. and J. Skolnick, Automated structure prediction of weakly homologous proteins on genomic scale. Proc. Natl. Acad. Sci. (USA), 2004. 101: p. 7594--7599.
2. Zhou, H. and J. Skolnick, Protein model quality assessment prediction by combining fragment comparisons and a consensus Cα contact potential. Proteins, 2007. 71: p. 1211--1218.
3. Zhang, Y. and J. Skolnick, SPICKER: a clustering approach to identify near-native protein fold. J. Comput Chem, 2004. 25 p. 865--871.
4. Zhou, H. and J. Skolnick, Ab initio protein structure prediction using chunk-TASSER. Biophys. J., 2007. 93: p. 1510--1518.
5. Zhou, H., et al., Analysis of TASSER based CASP7 protein structure prediction results. Proteins, 2007. 69(S8): p. 90--97.
6. Lee, S. and J. Skolnick, Benchmarking of TASSER_2.0: An improved protein structure prediction algorithm with more accurate predicted contact restraints. Biophys. J., 2008. 95: p. 1956--1964.
7. Rotkiewicz, P. and J. Skolnick, Fast procedure for reconstruction of full-atom protein models from reduced representations. Journal of Computational Chemistry, 2008. 29: p. 1460--1465.

TJ_Jiang

Tertiary Structure Prediction by a Combination of Threading and Fragment-based Assembly

T. Jiang, A. Wu, Y. Cao, L. Tian, Y. Hu, Z. Miao, L. Deng
Center for Computational and Systems Biology, Institute of Biophysics,
Chinese Academy of Sciences
taijiao@moon.ibp.ac.cn

To predict structures for all 57 human-expert targets in CASP8, we present a hybrid method based on combination of threading and fragment-based assembly. Briefly, for a target sequence, the method first attempted to find a structural template for the full-length protein by threading it to 91686 known structures of single-chain proteins that were derived from Protein Data Bank (PDB) [1] as of May 3, 2008, and then evaluated the quality of its structural templates. If a structural template of high quality was obtained, the structural template was used directly to build the structure for the target sequence by MODELLER [2].

Otherwise, the protein was parsed into domains by DOMAC [3] and then the structures of individual domains were modeled using the same procedure as for the full-length sequence. For those domains that could not find reliable templates, we used a fragment-based assembly strategy.

Our efforts were mainly focused on developing the threading algorithm and fragment-based assembly algorithm.

For threading, we developed a sequence-structure alignment scoring function by considering not only sequence profile and secondary structure information [4] but also local conformational energy of three amino acids [5]. The sequence-structure alignment score was normalized by percentage of matched sites and the reference score of self alignment. The best template was the template structure with highest Z-score value.

For fragment-based assembly strategy, we developed a novel and effective five-bead coarse-grained scoring function which includes an atom-atom contact potential, a hydrogen-binding term and a triplet local conformational energy [5]. Moreover, we optimized several factors to improve the performance of the fragment-based assembly strategy [6,7]. The most remarkable ones include requiring the structural compatibility of adjacent structural elements and choice of different number of structural templates based on their statistical distributions.

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. Sali A., and Blundell T.L., (1993) Comparative modeling by satisfaction of spatial restraints. *J Mol. Biol.*Cheng J. (2007) DOMAC: An Accurate, Hybrid Protein Domain Prediction Server. *Nucleic Acids Research*. 35, 354-356
3. Cheng J. (2007) DOMAC: An Accurate, Hybrid Protein Domain Prediction Server. *Nucleic Acids Research*. 35, 354-356.
4. Zhou H., and 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.
5. Yang,J.S., Chen,W.W., Skolnick,J., and Shakhnovich,E.I. (2007) All-Atom Ab initio Folding of a Diverse Set of proteins. *Structure.* 15, 53-63
6. Carol A.R., Charlie E.M.S., Kira M.S.M., and Baker D., (2004) Protein Structure Prediction by Using Rosetta. *Methods in Enzymology*. 383, 66-93.
7. Ozkan S.B., Wu G.A., Chodera J.D., Dill K.A., (2007) Protein folding by zipping and assembly. *PNAS,* 104(29), 11987-11992.

Tripos_08

ORCHESTRAR Homology Modeling

R.E Smith
Tripos, St.Louis

Introduction
ORCHESTRAR is a suite of tools following the iterative process for the homology modeling of proteins, with the underlying theme of a knowledge-based approach using the information in HOMSTRAD[1].

The major components of the package include the programs CHORAL[2], CODA[3], ANDANTE[4] and HARMONY3. These packages were used in conjunction with the FUGUE homology recognition program. The user is provided with an ensemble of structurally conserved regions extracted from superposed parent structures. Structurally variable regions are then modeled by any one of three programs that access different loop solutions. Side-chain placement is aided by the use of parent information. Sequence-structure alignment evaluation and model validation is then performed. Poorly modeled regions can then be reassessed.

Methodology
1. Homologous Structure/Family Recognition.
This is performed by the program FUGUE[5].

2. Core construction
CHORAL makes use of differential geometry and pattern recognition algorithms to identify structurally conserved sections of superposed parent structures. Environment specific substitution tables (ESSTs) are used to classify and filter which patterns likely to represent the core of the target.

3. Loop building
The two algorithms, FREAD and PETRA, are used to predict loop solutions. FREAD uses a fragment database constructed high resolution structures found in the HOMSTRAD database. FREAD uses ESSTs derived substitutions when the environment of a residue is its' backbone dihedral angles. PETRA constructs loop solutions from a set of eight phi-psi pairs. These phi-psi pairs were derived from the identification of high individual amino acid propensities (hot spots) located within six partitions of the Ramachandran plot. The CODA method then does a pair wise comparison of all FREAD and PETRA predictions.

4. Side Chain Placement
The program ANDANTE utilizes ESST information based on observed side chain chi angle conservation of a large number of families in the HOMSTRAD database. The parent-target residue substitution allows Andante to borrow entire high probability side-chain conformations or to restrict rotamer library solutions to specific chi bins.

5. Model validation/Error detection in sequence-structure alignment
HARMONY3 is used to locate errors that may have occurred in the sequence structure alignment that have been carried through the model building process.
A score for each modeled residue is calculated based on five components; the amino acid propensity score for the observed environment, observed amino acid distribution for that residue obtained from a PSI-BLAST search, a propensity score for a residue based on observed ESST scores. A composite substitution score from merged ESSTs. Finally, a term for local alignment flexibility is calculated. Low scoring regions are then re-examined for errors in the sequence structure alignment or modeling errors.

1. Stebbings L.A. & Mizuguchi K. (2004) HOMSTRAD: recent developments of the Homologous Protein Structure Alignment Database. Nucleic Acids Res 32, D203-7.
2. Montalvao R.W., Smith R.E., Lovell S.C. & Blundell T.L. (2005) CHORAL: a differential geometry approach to the prediction of the cores of protein structures. Bioinformatics 21, 3719-25.
3. Deane C.M. & Blundell T.L. (2001) CODA: a combined algorithm for predicting the structurally variable regions of protein models. Protein Sci 10, 599-612.
4. Richard E. Smith, Simon C. Lovell, David F. Burke, Rinaldo W. Montalvao, Tom L. Blundell, Andante: Reducing side-chain rotamer search space during comparative modeling using environment-specific substitution probabilities Bioinformatics 23, 1099-1105.
5. Shi J., Blundell T. L. & Mizuguchi K. (2001) FUGUE: sequence-structurehomology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 310, 243-57.1.

YASARA

Homology modeling with optimized ligand interactions, pH-dependent hydrogen bonding networks and high-resolution refinement

E. Krieger[1], R.L. Dunbrack Jr.[2], R.W.W. Hooft[3] and B.Krieger[4]

[1] - CMBI 260, NCMLS, Radboud University Nijmegen Medical Center, PO Box 9101, 6500 HB Nijmegen, the Netherlands

[2] - Institute for Cancer Research, Fox Chase Cancer Center, 333 Cottman Avenue, Philadelphia PA 1911, USA

[3] – Bruker AXS, Oostsingel 209, 2612 HL Delft, the Netherlands

[4] – YASARA Biosciences, Neue-Welt-Hoehe 13/b, 8042 Graz, Austria

elmar@cmbi.ru.nl

The 'YASARA Structure' server (*www.yasara.org/homologymodeling*) submitted predictions for those CASP8 targets that could be built reliably using known template structures. Since active site residues tend to be the most conserved ones, we speculated that the interaction with ligands is also conserved in a way that allows to benefit from the inclusion of template ligands in the homology modeling procedure, even without any knowledge about the target ligands.

All template ligands were therefore fully parameterized using AutoSMILES (bond typing, hydrogen addition, point charge assignment (combining a RESP library with AM1BCC), and GAFF force field parameters, see *www.yasara.org/autosmiles*), so that their presence could be considered during loop modeling and rotamer prediction.

Special emphasis was put on the hydrogen bonding network: we developed a new global optimizer, which first analyzes ligands to identify ambiguous functional groups. These are groups that can either be rotated (like Asn/Gln side-chain amides in proteins) or adopt different pH-dependent protonation states (like the His side-chain imidazole). Then an interaction graph is built for all the groups in the model, and the SCWRL algorithm[1] is employed to solve the hydrogen bonding network using dead-end elimination and graph reduction to biconnected components (*www.yasara.org/hbondnet*).

The final high-resolution refinement was performed by running molecular dynamics simulations with the YASARA force field, which is based on the self-parameterizing YAMBER force field[2], but includes additional knowledge-based dihedral-angle potentials (*www.yasara.org/kbpotentials*).

1. Canutescu AA, Shelenkov AA and Dunbrack RL Jr. (2003). A graph-theory algorithm for rapid protein side-chain prediction.
   Protein Sci. **12**, 2001-2014.
2. Krieger E, Darden T, Nabuurs S, Finkelstein A, Vriend G (2004). Making optimal use of empirical energy functions: force field parameterization in crystal space. *Proteins* **57**, 678-683.

Zhang
Zhang-Server

Automated structure prediction by the I-TASSER pipeline

Y. Zhang, Y. Li, S. Wu, A. Roy

Center for Bioinformatics and Department of Molecular Bioscience, University of Kansas, 2030 Becker Dr, Lawrence, KS 66047

yzhang@ku.edu

The procedures we used for our human (as "Zhang") and server (as "Zhang-Server") structural predictions in CASP8 are the essentially same, except for that the human prediction exploited the templates and the residue contact information in CASP8 Server Section while Zhang-Server used our in-house programs. Both predictions are fully automatic. One of the main purposes is to develop and benchmark the algorithms for large-scale and automatic structure predictions.

The pipeline of I-TASSER predictions includes four general steps: template identification, structure reassembly, atomic model construction, and final model selection.

Template identification. The target sequences are threaded through a non-redundant PDB structure library for identifying appropriate global-structure templates (for TMB targets) or local fragments (for FM targets). Threading is done by MUSTER[1], which uses an extended sequence profile-profile alignment algorithm with the alignment score assisted by secondary structure match, structural fragment profile, solvent accessibility, backbone torsion angle, and hydrophobic scoring matrix. For hard targets, additional templates identified by LOMETS[2], a local meta-threading server including FUGUE[3], HHSEARCH[4], PROSPECT[5], PPA[6] and SP3[7], are used. In human prediction, we include additionally the models generated by other groups in the Server Section into the template pool. Having more threading templates from the Server Section is the only source of differences of Zhang and Zhang-Server predictions.

Structure assembly. Continuous fragments excised from the threading templates are exploited to assemble full-length models[6; 8] with unaligned loop regions built by *ab initio* modeling[9]. The I-TASSER potential includes four components: (1) general knowledge-based statistics terms from the PDB (C-alpha/side-chain correlations[9], H-bond[10] and hydrophobicity[11]); (2) spatial restraints from threading templates[2]; (3) sequence-based C-alpha contact predictions by SVMSEQ[12]; (4) distance and contact map from segmental threading[13]. The last two energy terms are relatively new to the pipe-line we used in the previous experiment[14]. The structure assembly iteration includes two steps[6]. The first step of simulations starts from the threading templates. In the second step, the simulation starts from the cluster centroids generated by SPICKER[15] which clusters all the trajectories from the first step of simulations. Spatial restraints collected the PDB structures searched by TM-align[16] based on the cluster centroids are also incorporated in the I-TASSER simulations. The purpose of the iteration is to refine the local geometry as well as the global topology of the SPICKER centroids.

Atomic model construction. The SPICKER cluster centroids from I-TASSER are reduced models with each residue represented by its C-alpha and side-chain center. The full-atomic models are built by REMO[17]. REMO is a new protocol we developed for constructing full-atomic model from C-alpha traces by

optimizing the hydrogen-bond network. The basic backbone fragments (C, N, O) are matched from a secondary structure specified backbone isomers library which consists of 68,206 non-redundant isomers from high-resolution PDB structures. The driving force in the REMO simulations includes H-bonding, clash/break-amendment, I-TASSER restraints, and CHARMM22 potential.

Model selection. The reduced models from I-TASSER simulations are ranked based on the structure density of SPICKER clusters[15]. For each of the reduced models, the atomic models from REMO simulations are selected based on an empirical scoring function which is equal to the number of H-bonds plus the TM-score[18] of the model with the SPICKER cluster centroid and the average TM-score of the model with the initial templates (for easy targets only). The weights of the empirical score have been trained in benchmark tests and the highest scoring models are finally submitted.

1. Wu, S. T. & Zhang, Y. (2008). MUSTER: Improving Protein Sequence Profile-Profile Alignments by Using Multiple Sources of Structure Information. Proteins, 10.1002/prot.21945
2. Wu, S. T. & Zhang, Y. (2007). LOMETS: A local meta-threading-server for protein structure prediction. Nucl Acids Res 35, 3375-3382.
3. 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.
4. Soding, J. (2005). Protein homology detection by HMM-HMM comparison. Bioinformatics 21, 951-60.
5. Xu, Y. & Xu, D. (2000). Protein threading using PROSPECT: design and evaluation. Proteins 40, 343-54.
6. Wu, S., Skolnick, J. & Zhang, Y. (2007). Ab initio modeling of small proteins by iterative TASSER simulations. BMC Biol 5, 17.
7. Zhou, H. & Zhou, Y. (2005). Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins 58, 321-8.
8. Zhang, Y. & Skolnick, J. (2004). Automated structure prediction of weakly homologous proteins on a genomic scale. Proc Natl Acad Sci U S A 101, 7594-7599.
9. Zhang, Y., Kolinski, A. & Skolnick, J. (2003). TOUCHSTONE II: A new approach to ab initio protein structure prediction. Biophys J 85, 1145-1164.
10. Zhang, Y., Hubner, I., Arakaki, A., Shakhnovich, E. & Skolnick, J. (2006). On the origin and completeness of highly likely single domain protein structures Proc Natl Acad Sci U S A 103, 2605-10.
11. Chen, H. & Zhou, H. X. (2005). Prediction of solvent accessibility and sites of deleterious mutations from protein sequence. Nucleic Acids Res 33, 3193-9.
12. Wu, S. & Zhang, Y. (2008). A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. Bioinformatics 24, 924-31.
13. Wu, S. T. & Zhang, Y. (2008). Identifying protein sub-structural templates by segmental threading. submitted.
14. Zhang, Y. (2006). Invited talk given at CASP7 conference, November 26-30, 2006, Asilomar Conference Center, Pacific Grove, CA.
15. Zhang, Y. & Skolnick, J. (2004). SPICKER: A clustering approach to identify near-native protein folds. J Comput Chem 25, 865-71.
16. Zhang, Y. & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 33, 2302-2309.
17. Li, Y. Q. & Zhang, Y. (2008). REMO: A New Protocol to Generate Full Atomic Protein Models from C-alpha Traces by Optimizing Backbone Hydrogen-Bonding Network. submitted.
18. Zhang, Y. & Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. Proteins 57, 702-710.

Zhou-SPARKS

Machine Learning as Part of the Solution to the Protein Structure Problem

E. Faraggi, B. Xue, Y. Yang, and Y. Zhou

Center for Computational Biology and Bioinformatics, Indiana University School of Medicine and Indiana University School of Informatics, Indiana University-Purdue University, Indianapolis, Indiana 46202, USA

yqzhou@iupui.edu

The problem of determining protein structure from sequence information has been at the forefront of modern biophysics. The challenges associated with this system include very large dimensionality and many non-linear interactions. Machine learners combined with extraction of relevant physical features

enable the good quality prediction of various one dimensional properties such as dihedral angles, secondary structure, accessible surface area, contact numbers, etc. In addition they are beginning to become useful as contact map predictors. We present two recent improvements[1,2] to the prediction of the dihedral angles and accessible surface area using a coordinate transformation and a new type of a machine learner that uses guided weights to better approximate interactions between the amino-acids in sequence space. These tools were incorporated into our current research to find and use those influences which determine the structure of proteins.

1. Xue, B., Dor, O., Faraggi, E. and Zhou Y. (2008). Real-value prediction of backbone torsion angles. *Proteins* **72**, 427-433.
2. Faraggi, E., Xue, B. and Zhou, Y. (2008) Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two-layer neural network. *Proteins In press*.

Zico

Hierarchy of General Linear Models for Selecting and Ranking the Best Predicted Protein Structures

H.Z. Girgis[1,2], D. Fischer[1]
[1] - *The State University of New York at Buffalo,*
[2] - *The Johns Hopkins University*
hzgirgis@buffalo.edu

Protein structure prediction methods generate a large number of the predicted structures. Predicting how similar a predicted structure to the unknown experimentally determined structure is an important open problem. Structural biologists have developed several model quality assessment programs to determine the quality of the predicted models. Model quality assessment programs suffer from two limitations: *(i*) the rank-one structure is not necessarily the best predicted structure, in other words, the best predicted structure could be ranked as the 10th structure *(ii*) no single assessment method can correctly rank the predicted structures for all target proteins. Therefore; a model quality assessment method that is based on a consensus of other model quality assessment methods is likely to perform better. We have designed an energy function based on a hierarchy of general linear model. The hierarchical model consists of three levels. At the first level, the model quality assessment program DFire[7] selects the top 80 predicted structures to eliminate low quality structures. At the second level, a linear classifier separates the top 40 structures from the lower 40 structures. At the third level a linear regression model ranks the top 40 predicted structures. My method is based on a consensus of five model quality assessment programs. Next we give the details of my machine learning based model quality assessment program. Figure 1 outlines the system.

Data: the available data consists of the predicted 3d-models submitted to CASP6, CASP7, and CASP8. we used CASP6 data in training, servers' predictions in CASP7 in validation, and CASP7 humans' predictions and CASP8 servers' predictions in testing. We consider all of the five 3d-models submitted by each group not only the rank-one 3d-models. All data are preprocessed as the following: *(i)* select models that are at least 85% complete *(ii)* select full atoms models *(iii)* run the Modeller[2] program on the selections, then remove models whose MaxSub[3] score with the original model is less than 0.85. During training, target proteins whose best predicted model has a MaxSub[3] score less than 0.3 are excluded from the training set. In addition, all 3d-models which have MaxSub scores less than 0.1 are also removed from the training set.

Target and Features: the learning model predicts the MaxSub rank. The MaxSub score is the three dimensional similarity between two protein structures. Our learning algorithm uses only five features which are the ranks assigned by the following five model quality assessment programs: ProQ[4], Prosa-pair[5], ModCheck[6], DFire[7], and the 3dSim. The 3dSim score is the average of the MaxSub scores between a 3d-model and the other 3d-models predicted for the same target protein provided that the MaxSub similarity score between the two 3d-models is greater than 0.4. We use two rank-based normalization methods at the second and the third levels to standardize the data. In both cases, the normalization methods used are target-wise normalization. In other words, we standardize the values of each feature with regard to the structures predicted to the same target protein only.

First Level: the model quality assessment program DFire[7] selects the top 80 structures. We assume that the predicted structures whose DFire's ranks are greater than 80 are noisy structures. In addition, the goal of a model quality assessment program is to help the structural biologist to find a few good predicted structures, usually five 3d-models. Therefore, selecting 80 predicted structures are reasonable since the structural

biologists are interested in the best 3d-model or the best five 3d-models.

Second Level: the classifier learns to separate two classes: *(i)* the class of the top 40 ranks (1-40) and *(ii)* the class of the lower 40 ranks (41-80). We choose to make each class contains 40 structures to make sure that the classifier is not biased to any of the two classes. We train the linear classifier on the set $D_1 = \{(x^1,a_1),...,(x^q,a_q)\}$. Where q = 80 x *n*, such that *n* is the number of the proteins in the training set. Input x^i is a 5-dimensional vector representing the five features of the structure i, $x^i = \{x_1,...,x_5\}$ where $x_j = 1$ if the j^{th} feature's rank is below 40 and is -1 o.w. and a_i is the target to be learnt such that a_i is 1 if the MaxSub's

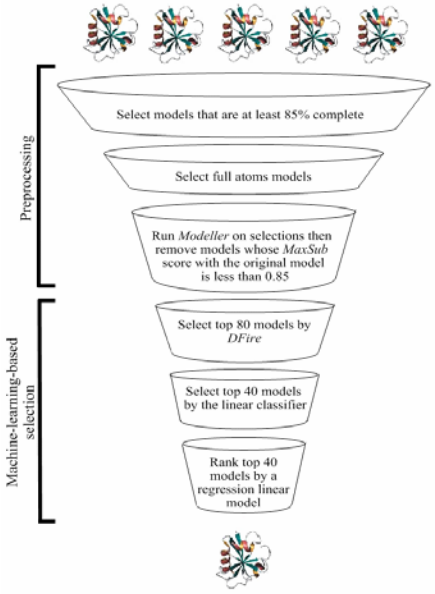


Figure 1: an overall view of the hierarchical model

rank is below 40 and -1 o.w. For example, the pair ([1 1 -1 1 1], 1) means that the 1^{st}, 2^{nd}, 4^{th}, and 5^{th} features rank the structure within the top 40 ranks, and the 3^{rd} feature ranks the structure within the lower 40 ranks. MaxSub ranks the predicted structure within the top 40 ranks.

Third Level: the linear classifier in the previous stage selects 40 models to pass to the third and final stage. A regression linear model is trained to predict the MaxSub's ranks of the 3d-models. We train the linear regression model on set $D_2 = \{(x^1,a_1),...,(x^k,a_k)\}$. Where k = 40 x *n*, such that *n* is the number of proteins in the training set. Input x^i is a 5-dimensional vector representing the five features of a structure i. $x^i = \{x_1,...,x_5\}$, where x_j is the model's rank assigned by the j^{th} feature, and $x_j \in \{1,2,..,39,40\}$. The output a_j is the rank assigned by MaxSub, and $a_j \in \{1,2,..,39,40\}$. For example, the pair ([1 5 7 3 4], 3) means that the five features rank the structure as the 1^{st}, the 5^{th}, the 7^{th}, the 3^{rd}, and the 4^{th} structure, and MaxSub puts the structure on the 3^{rd} rank. We have set the two thresholds to 80 and 40 to participate in CASP8 based on the number of the predicted structures per target protein in CASP6 and CASP7. However, these two thresholds should be increased when the model quality assessment program is applied to computational methods that produce a larger number of predicted structures and vice versa.

Results: We have evaluated the hierarchical model on a set consists of the CASP7 structures predicted by the human predictors. Our model quality assessment program outperforms the best human predictor by 2.9% based on the MaxSub scores of the rank-one structures. Our method outperforms the best performing component of its five components model quality assessment programs by 7.7%.

1. Hill,T. & Lewicki,P. (2007). Statistics Methods and Applications. StatSoft, Tulsa, OK.
2. Sali,A. & Blundell,T. (1993). Comparative protein modeling by satisfaction of spatial restraints. Journal of Molecular Biology, 234, 779–815.
3. Siew,N., Elofsson,A., Rychlewski,L., & Fischer,D. (2000). Maxsub: an automated measure for the assessment of protein structure prediction quality. Bioinformatics, 16, 776–785.
4. Wallner,B. & Elofsson,A. (2003). Can correct protein models be identified? Protein Science, 12, 1073–1086.
5. Sippl,M.J. (1995). Knowledge-based potentials for proteins. Curr. Opin. Struct. Biol., 5, 229–235.
6. Jones,D. (1999). Genthreader: An efficient and reliable protein fold recognition method for genomic sequences. Journal of Molecular Biology, 287, 797–815.
7. Zhou,H. and Zhou,Y. (2002). Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Science, 11, 2714–2726.

ZicoFullSTP

On-line Hierarchy of General Linear Models for Selecting and Ranking the Best Predicted Protein Structures

H.Z. Girgis[1,2], D. Fischer[1] J.J. Corso[1]
[1] - *The State University of New York at Buffalo,*
[2] - *The Johns Hopkins University*
hzgirgis@buffalo.edu

Protein structure prediction methods generate a large number of the predicted structures. Predicting how similar a predicted structure to the unknown experimentally determined structure is an important open problem. Structural biologists have developed several model quality assessment programs to determine the quality of the predicted models. Model quality assessment programs suffer from two limitations: *(i*) the rank-one structure is not necessarily the best predicted structure, in other words, the best predicted structure could be ranked as the 10th structure *(ii*) no single assessment method can correctly rank the predicted structures for all target proteins. Therefore; a model quality assessment method that is based on a consensus of other model quality assessment methods is likely to perform better. We have devised the STPdata algorithm. We have applied it to build an on-line "custom-trained" hierarchy of general linear models to select and rank the best predicted structures. By "custom-trained", we mean for each target protein the STPdata algorithm trains a unique model on data related to the input target protein. In CASP8, the STPdata algorithm has trained 128 hierarchical models for the 128 target proteins. Our method is based on a consensus of five model quality assessment programs. Next we give the details of our method.

Data: the available data consists of the predicted 3d-models submitted to CASP6, CASP7, and CASP8. We used CASP6 data in training, servers' predictions in CASP7 in validation, and CASP7 humans' predictions and CASP8 servers' predictions in testing. We consider all of the five 3d-models submitted by each group not only the rank-one 3d-models. All data are preprocessed as the following: *(i)* select models that are at least 85% complete *(ii)* select full atoms models *(iii)* run the Modeller[3] program on the selections, then remove models whose MaxSub[4] score with the original model is less than 0.85. During training, target proteins whose best predicted model has a MaxSub[4] score less than 0.3 are excluded from the training set. In addition, all 3d-models which have MaxSub scores less than 0.1 are also removed from the training set.

Target and Features: our learning model predicts the MaxSub rank. The MaxSub score is the three dimensional similarity between two protein structures. Our learning algorithm uses only five features which are the ranks assigned by the following five model quality assessment programs: ProQ[5], Prosa-pair[6], ModCheck[7], DFire[8], and the 3dSim. The 3dSim score is the average of the MaxSub scores between a model and the other models predicted for the same target protein provided that the MaxSub similarity score between the two models is greater than 0.4. We use the 0-1 normalization method and two rank-based normalization methods at different stages to standardize the data. In both cases, the normalization methods we use are target-wise normalization. In other words, we standardize the values of each feature with regard to the models predicted to the same target protein only.

The STP: Sample-**T**rain-**P**redict[1] **algorithm:** We apply the STP algorithm when the available data have two main properties. First, the available training (labeled) data is constantly growing. For example, the protein structure bank is increasing in size on a weekly basis. Second, the data is intrinsically clustered based on similarity in sequence, structure or function (each cluster has high-level semantic meaning). For instance, a set of predicted structures to the same target protein is viewed as a cluster in our current work. We describe the application of the STPdata algorithm in this abstract. The STP algorithm does its prediction in a batch mode i.e. it takes a cluster of data of unknown target values as its input and outputs the results in a batch mode as well. The STP algorithm has three stages as shown in figure 1: (i) Sample: select a subset of the training data based on the similarity to the unlabeled data; we use the distribution of the 3dSim scores as the similarity measures between the input (test) cluster and the clusters stored in the database (ii) Train: train a hierarchy of general linear models on the sampled data (iii) Predict: use the trained hierarchical model to select and rank the best predicted structures. We regard the STPdata algorithm as a method to build a custom-trained expert designed specifically to the input target protein predictions.

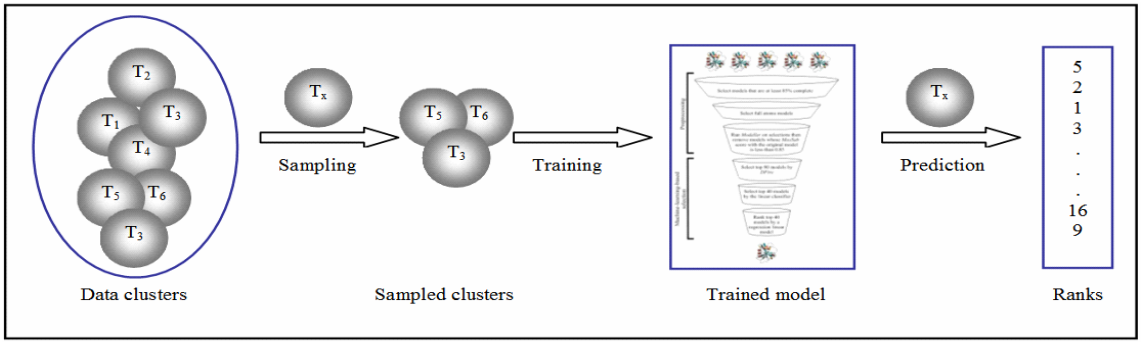


Figure 1: The STP algorithm has three stages: Sample, Train, and Predict

STPdata Sampling: STPdata considers the features' scores of the structures predicted to the same target protein as one cluster. We represent each cluster by two centers of the bimodal distribution of the 3dSim scores and the percentages of the predicted structures that belong to each mode. We obtain the two centers by applying the k-means clustering algorithm with initial centers 0.0 and 1.0 to the 3dSim scores of each cluster. For example, the vector [0.8 0.1 0.6 0.4] means that the k-means algorithm found two centers at 0.8 and 0.1, and 60% of the predicted structure are clustered around the 0.8 center and the other 40% of the predicted structures are clustered around the 0.1 center. We represent the input cluster in a similar fashion. Then, we apply the k-nearest algorithm to the clusters representations to find the nearest 22 clusters to the input cluster. We have decided to use 22 clusters based on the experimental results on the training and the validation sets.

STPdata Training: in this stage the algorithm trains a custom-made hierarchy of general linear models $(GLM)^2$ specifically to the input cluster. The hierarchy of general linear model consists of three levels. At the first level, DFire selects the top 80 predicted structures to eliminate low quality structures. At the second level, a linear classifier separates the top 40 structures from the lower 40 structures. At the third level a linear regression model ranks the top 40 predicted structures.

STPdata Prediction: once the on-line custom-trained hierarchy of GLM's is trained on the related clusters to the input cluster, the STPdata algorithm outputs the predicted ranks of the top 40 structures.

Results: We have evaluated our method on a set consists of the CASP7 structures predicted by the human predictors. Our method outperforms the best human predictor by 3% based on the MaxSub scores of the rank-one structures. Our method outperforms the best performing component of its five components model quality assessment programs by 10%.

1. Girgis,H.Z. & Corso,J.J. (2008). Stp: the sample-train-predict algorithm and its application to protein structure meta-selection. Technical Report 2008-16, The State University of New York at Buffalo.
2. Hill,T. & Lewicki,P. (2007). *Statistics Methods and Applications*. StatSoft, Tulsa, OK.
3. Sali,A. & Blundell,T. (1993). Comparative protein modeling by satisfaction of spatial restraints. *Journal of Molecular Biology*, **234**, 779–815.
4. Siew,N., Elofsson,A., Rychlewski,L., & Fischer,D. (2000). Maxsub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics*, **16**, 776–785.
5. Wallner,B. & Elofsson,A. (2003). Can correct protein models be identified? *Protein Science*, **12**, 1073–1086.
6. Sippl,M.J. (1995). Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.*, **5**, 229–235.
7. Jones,D. (1999). Genthreader: An efficient and reliable protein fold recognition method for genomic sequences. *Journal of Molecular Biology*, **287**, 797–815.
8. Zhou,H. and Zhou,Y. (2002). Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Science*, **11**, 2714–2726.

Application of Solid-State NMR Restraint Potentials in Membrane Protein Modeling

J. Lee[1], J. Chen[2], C. L. Brooks III[3] and W. Im[4]
[1] - *Department of Computational Sciences, Korea Institute for Advanced Study, Korea*
[2] - *Department of Biochemistry, The Kansas State University, USA*
[3] - *Department of Chemistry, The University of Michigan, USA*
[4] - *Department of Molecular Biosciences and Center for Bioinformatics, The University of Kansas, USA*

We have developed a set of orientational restraint potentials for solid-state NMR observables including ^{15}N chemical shift and ^{15}N-^{1}H dipolar coupling. A series of assessments show that the calculated restraint forces are numerically accurate. Torsion angle molecular dynamics simulations with available experimental ^{15}N chemical shift and ^{15}N-^{1}H dipolar coupling as target values have been performed to determine orientational information of four membrane proteins and to model the structures in oligomer states. The results suggest that incorporation of the orientational restraint potentials into molecular dynamics provides an efficient means to the determination of the structures that optimally satisfy the experimental observables without an extensive geometrical search.

Author Index

I

J

K

L

M

N

O

P

R

S

T

U

V

W

X

Y

Z

Table of contents