Prediction of the three dimensional structure of proteins using the electrostatic screening model

F. Avbelj

The three-dimensional structure of proteins is predicted from their sequence alone. The method is based on the electrostatic screening model for the stability of the protein main-chain conformation (1,2). According to the electrostatic screening model the stability of a main-chain conformational state of an amino acid in a protein depends primarily on the strengths of local and short-range nonlocal main-chain electrostatic interactions. The strength of local and nonlocal electrostatic interactions is related to the electrostatic screening with solvent and protein groups. The local main-chain electrostatic interactions are primarily due to the interaction of the main-chain CO and NH groups within an amino acid. The nonlocal main-chain electrostatic interactions are predominantly due to the main-chain hydrogen bonding. The free energy of a protein as a function of its conformation is obtained from the potentials of mean force analysis of high resolution x-ray protein structures (3). The free energy function is simple and contains only 44 fitted coefficients (4). The minimization of the free energy is performed by the torsion (5) and cartesian space (6) Monte Carlo procedure. In the first phase of the minimization procedure only the short-range interactions are activated. Short-range interactions are interactions between amino acids less than four residues apart in the sequence. The majority of $\alpha$-helices and $\beta$-strands are formed during the initial phase. The long-range interactions are activated in later phases which is causing the condensation of $\alpha$-helices and $\beta$-strands into super-secondary and the larger compact structures. Long-range interactions are interactions between the amino acids distant in the sequence. References: 1. F. Avbelj and J. Moult, Biochemistry, 34, 755-764 (1995). 2. F. Avbelj and L. Fele, J. Mol. Biol., 279, 665-684 (1998). 3. F. Avbelj, Biochemistry, 31, 6290-6297 (1992). 4. F. Avbelj and L. Fele, Proteins: Struc., Funct., Genet., 31, 74-96 (1998). 5. F. Avbelj and J. Moult, Proteins: Struc., Funct., Genet., 23, 129-141, (1995). 6. F. Avbelj in preparation.

Folding proteins with distance geometry and predicted inter-residue distances

Enoch Huang, Ram Samudrala, and Jay Ponder

Our method tackles the sub-problems of conformational search and scoring separately. To sample conformational space, we use the technique of distance geometry. By specifying only the most generic of constraints, we repeatedly generate folds with the hopes of generating native-like folds after a reasonable number of trials. Our method is tailored for small targets exhibiting a propensity to form helices as predicted by the PHD method (Rost B, Sander C 1993, J Mol Biol 232: 584-599) and/or the consensus described by the Jpred server (http://circinus.ebi.ac.uk:8081/). These helices were rigidly assembled into tertiary folds by enforcing generic inter-helical distances designed to enforce compactness (typically 5-11 or 5-15 Angstroms, depending on the size of the target). These inter-helical distances were defined as those between designated

CA atoms, one on each helix. The designated CA atoms were from those residues closest to the center of each predicted helix that were also predicted by PHD with high confidence (SUB acc) to be buried. Generic distances between each predicted helix was thus assigned. Prior to distance geometry calculations, all residues other than Gly and Pro were converted to Ala . All calculations were performed with the program distgeom from the TINKER suite (http://dasher.wustl.edu/tinker/). Each generated structure was refined via 10000 steps of simulated annealing against a set of penalty functions which enforce local geometry, chirality, excluded volume, input distance restraints, and torsion restraints. Over the residues predicted to be helical, two types of restraints were specified: a virtual torsion angle defined by four consecutive CA atoms, constrained to be between 40 and 60 degrees, and intra-helical CA-CA distances culled from a canonical helix. Structures that failed to anneal to right-handed helices were discarded. After on the order of 1000 structures were generated, side-chains were modeled using the program SegMod (Levitt M. 1992, J Mol Biol 226: 507-33) followed by minimization with ENCAD (Levitt M, Hirshberg M, Sharon R, Daggett V 1995, Comp. Phys. Commun., 91: 215-231). These folds were then scored with a hybrid scoring function, i.e. a linear combination of three functions (Shell: Park BH, Huang ES, Levitt M. 1997, J Mol Biol 266:831-46; RAPDF: Samudrala R, Moult J.1998, J Mol Biol 275:895-916; HCF: Samudrala R Xia Y, Levitt M, Huang ES. Pac Symp Biocomput 1999). Consensus distance geometry was used to build a few final models (Huang ES, Samudrala R, Ponder JW 1998, Protein Sci 7: 1998-2003), varying over different consensus conditions. Models were refined by fitting with a 4-state off-lattice model (Park BH, Levitt M. J Mol Biol 1995,249:493-507) and minimized with ENCAD. Final selection of the models was typically done with the assistance of the all-atom function RAPDF, the hydrophobic fitness function (Huang ES, Subbiah S, Levitt M. 1995, J Mol Biol 252: 709-20), and visual inspection. In some cases the best scoring fold with respect to the hybrid function was also submitted.

Improving the sensitivity of Hidden Markov Models for protein fold recognition

Jeanette Hargbo, Bj?n Larsson and Arne Elofsson

We have developed a rigorous benchmark for protein fold recognitions where representatives for all proteins of known structure are matched against each other. Using this benchmark we have compared the performance of automatically created hidden Markov models with standard sequence search methods, as well as linking methods. In our benchmark it was shown that a HMM that used no multiple sequence information and no predicted secondary structures correctly identified the fold of a protein in 10~\% of the cases. Including multiple sequence information increased this number to 16~\% and when also predicted secondary structure information was included the fold was correctly identified 20~\% of the times. Further we have studied other methods to increase the sensitivity of HMMs for fold recognition. We have studied the use of different substitution matrixes, gap-penalties, linking methods as well as different ways to construct the HMMs.

Comparative Modeling using SCWRL, a tool for choosing sidechain rotamers.

Dr. Michael J. Bower

Three targets were modeled: T0050, T0084, and T0085. T0050 had very high homology to the structure 1BE1, which was used as a scaffold. Residues were replaced on the 1BE1 backbone using SCWRL2.1 (Bower et al, J. Mol. Biol., 1997, 267, 1268-1282) (see http://www.cmpharm.ucsf.edu/~bower/scwrl/scwrl.html), which uses a backbone-dependent rotamer library and steric repulsions to place sidechains. SCWRL was allowed to place residues freely, or constrained to keep conserved residues in the positions in the scaffold. The second method, which keeps conserved residues in place, was used in the final model. Some steric clashes were relieved manually. A known problem is the backbone conformation of Pro 42 and the sidechain conformation of Val60. For T0084, a sequence search of the database of known structures revealed sequence elements similar to the leucine zipper structure of 1GCL, a synthetic peptide. Other similarities in the database of known structures include amphipathic helices, supporting the hypothesis that this is a designed four-helix bundle structure. 1GCL was used as a scaffold for the model, by matching and extending the supercoil structure. The four extended helices which resulted were used as the backbone structure for building sidechains. A sequence alignment to 1GCL was created which put the similar sidechains in the correct register with respect to the supercoil, and then sidechains were added. This was done with SCWRL. One strand of the four helix bundle was reported as the final prediction. For T0085, sequence searches indicated that the N-terminal half of this protein bore homology to the four-heme structures 2CY3, 2CDU, 2CYM, etc. The C-terminal half had homology to single-heme proteins like 1YEA. 2CYM and 1YEA were chosen as parent structures for the N- and C-terminal halves of the target, respectively. They were joined with a linker residue to form a two-domain scaffold. Very crude insertions and deletions were made to create the starting scaffold. Conserved residues were held in place, while non-conserved sidechains were built with SCWRL2.1. Sidechains in insertion and deletions were also free to move. Five heme molecules from the parent structures were used as steric clash checks. Some local minimization of the backbone in the region of insertions and deletions was accomplished with the program Flo96, followed by another round of SCWRL.

GLOBULAR PROTEIN SECONDARY STRUCTURE PREDICTION WITHOUT COMPUTER BY DOUBLET CODE(DOUC)METHOD

BORIS V. SHESTOPALOV

INTRO-DOUC-TION.The method is manual. It takes only 30 minutes to perform manual single sequence pre-DOUC-tion of 300 residue protein. The basis of the method was published in 1990 (Shestopalov B.V. Prediction of protein secondary structure by doublet code method. Mol. Biol., Moscow, Engl. transl., 24/4, p.900-907). For the CASP3 the method has been modified.
-- DOUC-SCRIPTION. Coils, strands, helices consist of overlaps of structurons which consist of 2, 3, 5, residues and are encoded by residue pairs (i, i+1), (i, i+2), (i, i+4) respectively. Codon tables are obtained from analysis of residue pairs occurence in secondary structures. Codon distributions in a primary structure are placed in three lines under the structure. Usually codons of

diiferent structural types overlap in an amino acid sequence. Choice of codons in such cases is necessary. The choice is to exclude the least number of codons until the overlap disappear. Obtained codon distributions are used for prediction. If several variants of distributions are obtained the prediction of some regions may be ambiguous and such regions can not be predicted at this stage. The average prediction accuracy of this procedure, so called single sequence prediction (SSP), is limited up to 63% because only local interactions are considered. If one uses similar sequences with such similar secondary structures predicted which contain as much as possible information about native secondary structure, the average secondary structure from their alignment may be nearer to the experimental one up to 5-10% and ambiguities are excluded. This is version of so called multiple sequence prediction (MSP) used here. ____ DOUC-TAILS. The codons are classified as strong and weak ones. A residue pair is strong (weak) codon if probability of respective structure for this pair is more (equal) than probability of total of two other structures. The probability is calculated from an occurence of the pair in a secondary structure database using the reverse binomial distribution (2P-1=0.999). The codon choice is performed firstly between strong codons. Then weak codons are considered. The secondary structure database was constructed firstly from primary and secondary structures of 257 proteins. Then secondary structure of these proteins was predicted by the code obtained from this database. Then new database was constructed from primary and secondary structures of correctly predicted regions and new code was obtained from this database. New code was used for new proteins secondary structure prediction and correctly predicted regions were added to the database and new code was obtained from the enlarged database and new proteins were predicted and new database was constructed... The DOUC-CODONS used for the CASP3 target prediction are remarked after the method. To this moment the DOUBLET CODE is ready up to 95-97%. Probably most of the weak codons with rare residues (W, C, H, M) may become strong ones or disappear. Therefore the version of DOUC method without differentiation between strong and weak codons is used also. Five models are used. Models 1, 2, 5 (3, 4) are MSP (SSP). M5 is the label for models developed and checked during the CASP3. Secondary structures for the alignments are selected from ones obtained by pre-DOUC-tions for sequences selected as mentioned before the model line. All the secondary structures which are similar to the target protein one not less than 60% are used for secondary structure alignment. Models 1 and 3 are ones without strong/weak differentiation, model 5 is tve variant the model 1 with the confidence level in all the range 0.00 - 1.00, that is not only values 0.00 and 1.00 are used as in the cases of models 1, 2, 3, 4. Model 5 is modified lastly on 05 September 1998 and this new version IS USED for targets 50, 70, 71, 72, 80, 81, 82, 83, 84, 85. The final version of models 1, 2 is used for all targets except targets 43, 52, 54 for which more restrictive version was used (similarity of the aligned secondary structuresto the target one is 70% then and 60% now) but this modification of M1 and M2 is not principal. ____ DOUC-CODONS-07.06.98. The database volume is 150000 amino acid residues. COIL. Strong codons: AD AG AP CG CN CP DC DD DG DH DK DN DP DS DT DW ED EG EP ES FP GA GC GD GE GG GH GK GL GM GN GP GQ GR GS GT GW HD HG HN HP HS IP KD KG KN KP KS LP MG MP ND NG NH NK NN NP NR NS NT PA PC PD PE PF PG PH PK PL PM PN PP PQ PR PS PT PW PY QG QP RG RP SD SG SH SK SN SP SQ SR SS ST TD TG TK TN TP TS VP WD WG WP YP. Weak codons: CD CH DQ DR EN HH HK KC MN MS NC NQ NW NY QD QN QS RN RS SC SW WC WN WS. STRAND. Strong codons: CI CV FC FF FI FL FT FV FW FY HV IC IF II IL IT IV IW IY LC LF LI LL LV LY MF MV TF TH TI TT TV TY VC VF VI VL VS VT VV VW VY WF WI WV WY YC YF YI YL YT YV YY. Weak codons: CC CF CL CM CT CW CY FM HC HF HH HI HW HY IM LW MC MI MW MY VH VM WC WH WL WW YH YW. HELIX. Strong codons: AA AE AK AL AM AQ AR EA EE EK EL EM EQ ER IL IM KA KE KQ LA LL LM LQ LR MA ME ML MR QA QD QE QK QQ QR RA RE RK RM RQ RR Weak codons: AH AW CM DR EH EW FL FM HE HH HM KM KR LI MH MI MK MM MQ MW QM RW WI WK WL WM WW YM. 127 COIL CODONS: 103 strong, 24 weak; 81 STRAND CODONS: 53/28; 68 HELIX CODONS: 40/28. 276 IN TOTALITY: 196/80.

GLOBULAR PROTEIN SECONDARY STRUCTURE PREDICTION WITHOUT COMPUTER BY DOUBLET CODE (DOUC) METHOD

BORIS V. SHESTOPALOV

INTRO-DOUC-TION.The method is manual. It takes only 30 minutes to perform manual single sequence pre-DOUC-tion of 300 residue protein. The basis of the method was published in 1990 (Shestopalov B.V. Prediction of protein secondary structure by doublet code method. Mol. Biol., Moscow, Engl. transl., 24/4, p.900-907). For the CASP3 the method has been modified. -- DOUC-SCRIPTION. Coils, strands, helices consist of overlaps of structurons which consist of 2, 3, 5, residues and are encoded by residue pairs (i, i+1), (i, i+2), (i, i+4) respectively. Codon tables are obtained from analysis of residue pairs occurence in secondary structures. Codon distributions in a primary structure are placed in three lines under the structure. Usually codons of diiferent structural types overlap in an amino acid sequence. Choice of codons in such cases is necessary. The choice is to exclude the least number of codons until the overlap disappear. Obtained codon distributions are used for prediction. If several variants of distributions are obtained the prediction of some regions may be ambiguous and such regions can not be predicted at this stage. The average prediction accuracy of this procedure, so called single sequence prediction (SSP), is limited up to 63% because only local interactions are considered. If one uses similar sequences with such similar secondary structures predicted which contain as much as possible information about native secondary structure, the average secondary structure from their alignment may be nearer to the experimental one up to 5-10% and ambiguities are excluded. This is version of so called multiple sequence prediction (MSP) used here. ____ DOUC-TAILS. The codons are classified as strong and weak ones. A residue pair is strong (weak) codon if probability of respective structure for this pair is more (equal) than probability of total of two other structures. The probability is calculated from an occurence of the pair in a secondary structure database using the reverse binomial distribution (2P-1=0.999). The codon choice is performed firstly between strong codons. Then weak codons are considered. The secondary structure database was constructed firstly from primary and secondary structures of 257 proteins. Then secondary structure of these proteins was predicted by the code obtained from this database. Then new database was constructed from primary and secondary structures of correctly predicted regions and new code was obtained from this database. New code was used for new proteins secondary structure prediction and correctly predicted regions were added to the database and new code was obtained from the enlarged database and new proteins were predicted and new database was constructed... The DOUC-CODONS used for the CASP3 target prediction are remarked after the method. To this moment the DOUBLET CODE is ready up to 95-97%. Probably most of the weak codons with rare residues (W, C, H, M) may become strong ones or disappear. Therefore the version of DOUC method without differentiation between strong and weak codons is used also. Five models are used. Models 1, 2, 5 (3, 4) are MSP (SSP). M5 is the label for models developed and checked during the CASP3. Secondary structures for the alignments are selected from ones obtained by pre-DOUC-tions for sequences selected as mentioned before the model line. All the secondary structures which are similar to the target protein one not less than 60% are used for secondary structure alignment. Models 1 and 3 are ones without strong/weak differentiation, model 5 is tve variant the model 1 with the confidence level in all the range 0.00 - 1.00, that is not only values 0.00 and 1.00 are used as in the cases of models 1, 2, 3, 4. Model 5 is modified lastly on 05 September 1998 and this new version IS USED for targets 50, 70, 71, 72, 80, 81, 82, 83, 84, 85. The final version of models 1, 2 is used for all targets except targets 43, 52, 54 for which more restrictive version was used (similarity of the aligned secondary structuresto the target one is 70% then and 60% now) but this modification of M1 and M2 is not principal. ____ DOUC-CODONS-07.06.98. The database volume is 150000 amino acid residues. COIL. Strong codons: AD AG AP CG CN CP DC DD DG DH DK DN DP DS DT DW ED EG EP ES FP GA GC GD GE GG GH GK GL GM GN GP GQ

GR GS GT GW HD HG HN HP HS IP KD KG KN KP KS LP MG MP ND NG NH NK NN NP NR NS NT PA PC PD PE PF PG PH PK PL PM PN PP PQ PR PS PT PW PY QG QP RG RP SD SG SH SK SN SP SQ SR SS ST TD TG TK TN TP TS VP WD WG WP YP. Weak codons: CD CH DQ DR EN HH HK KC MN MS NC NQ NW NY QD QN QS RN RS SC SW WC WN WS. STRAND. Strong codons: CI CV FC FF FI FL FT FV FW FY HV IC IF II IL IT IV IW IY LC LF LI LL LV LY MF MV TF TH TI TT TV TY VC VF VI VL VS VT VV VW VY WF WI WV WY YC YF YI YL YT YV YY. Weak codons: CC CF CL CM CT CW CY FM HC HF HH HI HW HY IM LW MC MI MW MY VH VM WC WH WL WW YH YW. HELIX. Strong codons: AA AE AK AL AM AQ AR EA EE EK EL EM EQ ER IL IM KA KE KQ LA LL LM LQ LR MA ME ML MR QA QD QE QK QQ QR RA RE RK RM RQ RR Weak codons: AH AW CM DR EH EW FL FM HE HH HM KM KR LI MH MI MK MM MQ MW QM RW WI WK WL WM WW YM. 127 COIL CODONS: 103 strong, 24 weak; 81 STRAND CODONS: 53/28; 68 HELIX CODONS: 40/28. 276 IN TOTALITY: 196/80.

Phylogenetic Approach to Detecting Tertiary Contacts

William J. Bruno, Aaron L. Halpern, Gerhard Hummer, Martijn A. Huynen

Our method can be summarized as evolutionary covariation analysis (also known as correlated mutations). We analyze an alignment, taking phylogeny into account, and look for significant non-independence of columns, beyond what can be explained by phylogeny (i.e., evolution). The first step is basically a generalization of the analysis in [Bruno, Mol. Biol. Evol. 1996] to considering pairs of sites. This involves a maximum-likelihood model of position-specific residues frequencies, estimated on the evolutionary tree. The model used ignores the genetic code; it is a model in which the choice of amino acid in a replacement event is independent of which amino acid occupied the site previously. Selection constraints are assumed to be constant over the tree. The model is reversible, and we assume that the most recent common ancestor of all sequences was sampled from the equilibrium distribution. Independence of sites is assumed in the calculation, and we calculate a 20 by 20 matrix of substitution events for each pair of sites. If the independence assumption is correct, these matrices should be consistent with row-column independence. The next step is a permutation test similar to (but different from) that of [Korber et al, PNAS USA, 1993] and others. Our test statistic comes from the hypergeometric distribution, and we use the standard statistical methods for correcting for multiple tests. To try to avoid being mislead by adaptive evolution (which our model does not include), we also eliminate columns that strongly correlate with the evolutionary clades (subtrees of the tree constructed from our alignment). We also checked for correlation with two other variables (luminous organisms, and taxonomic classification of organism), but no sites correlated significantly with either. Our alignment consisted of 23 of the closest homologs we could find to the target. The permutation test resulted in 13 putative contacts, each with 80% or higher probability after correcting for multiple tests. The adaptive evolution screen removed 4 columns (31, 97, 107, 219) responsible for 8 putative contacts, six of which were between members of this set. Our prediction consists of the remaining 5 contacts, which range from 80% to 97% probability. Our probability values represent the probability that mutations are non-independent; we have not attempted to correct for the fact that there may be other, non-phylogenetic causes of non-independence besides structural contact. We also have made no attempt to restrict to certain types of non-independence (e.g., salt-bridge conservation). All deviations from independence are treated equally. Conceivably, some deviations could arise from shortcomings in our model (such as neglecting the genetic code) or other approximations. However, previous control studies we have done suggest that these may not be the cause of false positives. Rather, we think any false positives probably represent real correlated evolution, caused by some kind of non-local interaction or constraint.

Homology modelling using key residues.

Alex Bateman

Given a basic homology between a known structure and a sequence the key residue method can be used to make accurate alignments which give a guide to the accessibility and position for each residue. The alignment can be used to make full atom coordinates for the sequence of unknown structure. During the last CASP meeting it was found that the alignment step was the most crucial to making a good homology model. The method used here comes in four parts: 1) Firstly a parent structure is chosen. In most cases this is the highest scoring match of the target with WU-blastp against the pdb100su database (12110 chains). pdb100su is a fasta file of sequences from pdb and from structures which are published but not yet available in pdb (kindly provided by Dr. Tim Hubbard). 2) A rough alignment is generated with clustalW (1.7) using the structural mask option. Secondary structure assignment was taken from DSSP. 3) Key residues are derived from the known parent structure. Key residues are those that mainly determine the conformation of the protein. They are mainly involved in packing, hydrogen bonding and/or have the ability to take up unusual torsion angles. The alignment is altered manually to maximise the similarity at key residue sites while retaining a structurally plausible alignment. 4) Construct full atom model with modeller 4. For some predictions regions that are expected to differ in conformation to the known structure are given zero occupancy. Details of the can be found in: Bateman A, Chothia C. Nature Struct. Biol. 2:1068-1074(1995). Bateman A, Jouet M, MacFarlane J, Du J, Kenwrick S, Chothia C. EMBO J. 15:6050-6059(1996).

Model Building by Comparison: Improving Algorithms via Expert Knowledge

Paul A. Bates and Michael J.E. Sternberg

Nine models were constructed for the comparative modelling section of CASP3; T0047, T0048, T0050, T0055, T0058, T0060, T0069, T0076, and T0082. A tenth target was also constructed that was not initially classified as a potential comparative model target, T0068. Sequence identity between each target and the best possible parent(s) ranged between 12 and 64%. The modelling protocol followed is an extension to the work reported in CASP2 (Bates et al., 1998). The method is a mixture of automated algorithms and manual intervention. Manual intervention is required at the critical sequence alignment stages and the selection of parameters for various computer programs. Seven of the targets were constructed from single parent template and three from multiple. The reasons for such a high ratio of modelling from single parents only are: (1) Only one parents available (2) The sequence spread between parents, relative to the target, isn't appropriate for the level of expected model error; (a) Sequence spread too small between parents to be meaningful in three dimensions, particularly, relative orientations between equivalent secondary structure elements. (b) Sequence spread between parents too large, thus, the parent with the highest overall sequence identity is expected to reduce modelling errors. The question of when a single parent is sufficient for modelling will be addressed. Where multiple parents are needed they are superimposed automatically using a program based upon a pairwise superposition algorithm (Gerstein & Levitt, 1996). The superimposed co-ordinates are then separated into conserved secondary structure elements and conserved loops. The conserved elements for the model are then chosen from the ensemble by a modification of the self consistent mean field approach to gap closure (Koehl & Delarue, 1995). The remaining loops, all loops in the case of single parent, and all regions with incompatible backbone angles with the target sequence were modelled via database fragments searches. Three database were searched in the order: (1) Homologous/analogous structures. (2) Loop classification database (Olivia et al., 1997). (3) Nonredundant database, 300 protein chains (sequence similarity of less than 25%, R-factor <= 2.5). Fragments were selected automatically and were chosen on the basis of good sequence similarity with the target and how well the fragments fitted to the take off points both in terms of rms fit on the Ca atoms used for the take off points and the difference in C=O angles of the backbones between template and chosen fragment in the take off region as described in (Bates et al., 1997), a modification of the Jones and Thirup approach (Jones & Thirup, 1986). Manual intervention was needed if candidate fragments could not be found to cover a region. In these cases different take off points were selected and/or different sequence and geometry tolerances set. A number of loop conformations were selected for each gap and the best candidate was chosen via a second mean field run, using the conserved secondary structure elements found in the first run. The frequency of loop selection from each of the three types of database will be addressed. Side-chains are built by tracing the path of the old side-chain. The maximum number of bond lengths, angles and torsion angles are taken from the old side-chain that are compatible with the new side-chain. Additional internal co-ordinates to complete the side-chain are taken from the secondary structure dependant rotamer library (McGregor et al., 1987) as described in (Bates et al., 1997). After the replacement of all side-chains and the assignment of a single rotamer for each, this parent rotamer, plus rotamers from a side-chain rotamer library are built at each residue position. Each rotamer feels the average environment due to rotamers of other residues weighted by their respective probabilities. The conformational probability matrix is refined to give resultant probabilities for the rotamers of each residue (Koehl & Delarue, 1994). Energy parameters taken from (Lee & Subbiah, 1991). The highest probability rotamer was chosen. The selected rotamers were inspected to check that all rotamers look sensible. If not then the side-chain rotamer from just the first stage in the side-chain replacement mechanism is used. Only a few rotamers needed to revert back and these were usually polar or charged residues involved in conserved hydrogen bonding interactions. The reasons for the mean field not always finding the best rotamer is addressed. To remove the small number of steric clashes remaining in the models 100 steps of steepest descents energy minimization (unrestrained) were run, on all models, using the program CHARM (CHARM software version 3.3 (1992) Molecular Simulations Inc. 200 Fifth Avenue, Waltham, MA 02154.). We have submitted unrefined and refined coordinates, before and after energy minimization. The significance of performing unrestrained global minimization, irrespective of the expected modelling errors, is addressed. References Bates, P.A., Jackson, R.M. Sternberg, M.J.E. Model building by comparison: A combination of expert knowledge and computer automation. Proteins, Suppl. 1,59-67, 1997. Gerstein, M., Levitt, M. Using interative dynamic programming to obtain accurate pairwise and multiple alignments of protein structures. Fourth International Conference on Intelligent Systems for Molecular Biology. AAAI Press, Menlo Park, CA, USA, 59-67, 1996. Nine models were constructed for the comparative modelling section of CASP3; T0047, T0048, T0050, T0055, T0058, T0060, T0069, T0076, and T0082. A tenth target was also constructed that was not initially classified as a potential comparative model target, T0068. Sequence identity between each target and the best possible parent(s) ranged between 12 and 64%. The modelling protocol followed is an extension to the work reported in CASP2 (Bates et al., 1998). The method is a mixture of automated algorithms and manual intervention. Manual intervention is required at the critical sequence alignment stages and the selection of parameters for various computer programs. Seven of the targets were constructed from single parent template and three from multiple. The reasons for such a high ratio of modelling from single parents only are: (1) Only one parents available (2) The sequence spread between parents, relative to the target, isn't appropriate for the level of expected model error; (a) Sequence spread too small between parents to be meaningful in three dimensions, particularly, relative orientations between equivalent secondary structure elements. (b) Sequence spread between parents too large, thus, the parent with the highest overall sequence identity is expected to reduce modelling errors. The question of when a single parent is sufficient for modelling will be addressed. Where multiple parents are needed they are superimposed automatically using a program based upon a pairwise superposition algorithm (Gerstein & Levitt, 1996). The superimposed coordinates are then separated into conserved secondary structure elements and conserved loops. The conserved elements for the model are then chosen from the ensemble by a modification of the self consistent mean field approach to gap closure (Koehl & Delarue, 1995). The remaining loops, all loops in the case of single parent, and all regions with incompatible backbone angles with the target sequence were modelled via database fragments searches. Three database were searched in the order: (1) Homologous/analogous structures. (2) Loop classification database (Olivia et al., 1997). (3) Nonredundant database, 300 protein chains (sequence similarity of less than 25%, R-factor <= 2.5). Fragments were selected automatically and were chosen on the basis of good sequence similarity with the target and how well the fragments fitted to the take off points both in terms of rms fit on the Ca atoms used for the take off points and the difference in C=O angles of the backbones between template and chosen fragment in the take off region as described in (Bates et al., 1997), a modification of the Jones and Thirup approach (Jones & Thirup, 1986). Manual intervention was needed if candidate fragments could not be found to cover a region. In these cases different take off points were selected and/or different sequence and geometry tolerances set. A number of loop conformations were selected for each gap and the best candidate was chosen via a second mean field run, using the conserved secondary structure elements found in the first run. The frequency of loop selection from each of the three types of database will be addressed. Side-chains are built by tracing the path of the old side-chain. The maximum number of bond lengths, angles and torsion angles are taken from the old side-chain that are compatible with the new side-chain. Additional internal co-ordinates to complete the side-chain are taken from the secondary structure dependant rotamer library (McGregor et al., 1987) as described in (Bates et al., 1997). After the replacement of all side-chains and the assignment of a single rotamer for each, this parent rotamer, plus rotamers from a side-chain rotamer library are built at each residue position. Each rotamer feels the average environment due to rotamers of other residues weighted by their respective probabilities. The conformational probability matrix is refined to give resultant probabilities for the rotamers of each residue (Koehl & Delarue, 1994). Energy parameters taken from (Lee & Subbiah, 1991). The highest probability rotamer was chosen. The selected rotamers were inspected to check that all rotamers look sensible. If not then the side-chain rotamer from just the first stage in the side-chain replacement mechanism is used. Only a few rotamers needed to revert back and these were usually polar or charged residues involved in conserved hydrogen bonding interactions. The reasons for the mean field not always finding the best rotamer is addressed. To remove the small number of steric clashes remaining in the models 100 steps of steepest descents energy minimization (unrestrained) were run, on all models, using the program CHARM (CHARM software version 3.3 (1992) Molecular Simulations Inc. 200 Fifth Avenue, Waltham, MA 02154.). We have submitted unrefined and refined coordinates, before and after energy minimization. The significance of performing unrestrained global minimization, irrespective of the expected modelling errors, is addressed.

References Bates, P.A., Jackson, R.M. Sternberg, M.J.E. Model building by comparison: A combination of expert knowledge and computer automation. Proteins, Suppl. 1,59-67, 1997. Gerstein, M., Levitt, M. Using interative dynamic programming to obtain accurate pairwise and multiple alignments of protein structures. Fourth International Conference on Intelligent Systems for Molecular Biology. AAAI Press, Menlo Park, CA, USA, 59-67, 1996. Jones, T.A., Thirup, S. Using known substructures in protein model building and crystallography. EMBO J. 5, 819-823, 1986. Koehl, P., Delarue, M. Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformation entropy. J. Mol. Biol. 239, 249-275, 1994. Koehl, P., Delarue, M. A self consistent mean field approach to simultaneous gap closure and side-chain positioning in homology modelling. Nat. Struc. Biol. 2, 163-170, 1995. Lee C., Subbiah S. Prediction of protein side-chain conformation by packing optimization. J. Mol. Biol. 217, 373-388, 1991. McGregor, M.J., Islam, S.A., Sternberg, M.J.E. Analysis of the relationship between side-chain conformation and secondary structure in globular proteins. J. Mol. Biol. 198, 295-310, 1987. Olivia, B., Bates, P.A., Enrique, Q., Aviles, F.X., Sternberg, M.J.E. An automated classification of the structure of protein loops. J. Mol. Biol. 266,814-830, 1997.

Jones, T.A., Thirup, S. Using known substructures in protein model building and crystallography. EMBO J. 5, 819-823, 1986. Koehl, P., Delarue, M. Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformation entropy. J. Mol. Biol. 239, 249-275, 1994. Koehl, P., Delarue, M. A self consistent mean field approach to simultaneous gap closure and side-chain positioning in homology modelling. Nat. Struc. Biol. 2, 163-170, 1995. Lee C., Subbiah S. Prediction of protein side-chain conformation by packing optimization. J. Mol. Biol. 217, 373-388, 1991. McGregor, M.J., Islam, S.A., Sternberg, M.J.E. Analysis of the relationship

between side-chain conformation and secondary structure in globular proteins. J. Mol. Biol. 198, 295-310, 1987. Olivia, B., Bates, P.A., Enrique, Q., Aviles, F.X., Sternberg, M.J.E. An automated classification of the structure of protein loops. J. Mol. Biol. 266,814-830, 1997.

Model Building by Comparison: Improving Algorithms via Expert Knowledge

Paul A. Bates and Michael J.E. Sternberg

Nine models were constructed for the comparative modelling section of CASP3; T0047, T0048, T0050, T0055, T0058, T0060, T0069, T0076, and T0082. A tenth target was also constructed that was not initially classified as a potential comparative model target, T0068. Sequence identity between each target and the best possible parent(s) ranged between 12 and 64%. The modelling protocol followed is an extension to the work reported in CASP2 (Bates et al., 1998). The method is a mixture of automated algorithms and manual intervention. Manual intervention is required at the critical sequence alignment stages and the selection of parameters for various computer programs. Seven of the targets were constructed from single parent template and three from multiple. The reasons for such a high ratio of modelling from single parents only are: (1) Only one parents available (2) The sequence spread between parents, relative to the target, isn't appropriate for the level of expected model error; (a) Sequence spread too small between parents to be meaningful in three dimensions, particularly, relative orientations between equivalent secondary structure elements. (b) Sequence spread between parents too large, thus, the parent with the highest overall sequence identity is expected to reduce modelling errors. The question of when a single parent is sufficient for modelling will be addressed. Where multiple parents are needed they are superimposed automatically using a program based upon a pairwise superposition algorithm (Gerstein & Levitt, 1996). The superimposed co-ordinates are then separated into conserved secondary structure elements and conserved loops. The conserved elements for the model are then chosen from the ensemble by a modification of the self consistent mean field approach to gap closure (Koehl & Delarue, 1995). The remaining loops, all loops in the case of single parent, and all regions with incompatible backbone angles with the target sequence were modelled via database fragments searches. Three database were searched in the order: (1) Homologous/analogous structures. (2) Loop classification database (Olivia et al., 1997). (3) Nonredundant database, 300 protein chains (sequence similarity of less than 25%, R-factor <= 2.5). Fragments were selected automatically and were chosen on the basis of good sequence similarity with the target and how well the fragments fitted to the take off points both in terms of rms fit on the Ca atoms used for the take off points and the difference in C=O angles of the backbones between template and chosen fragment in the take off region as described in (Bates et al., 1997), a modification of the Jones and Thirup approach (Jones & Thirup, 1986). Manual intervention was needed if candidate fragments could not be found to cover a region. In these cases different take off points were selected and/or different sequence and geometry tolerances set. A number of loop conformations were selected for each gap and the best candidate was chosen via a second mean field run, using the conserved secondary structure elements found in the first run. The frequency of loop selection from each of the three types of database will be addressed. Side-chains are built by tracing the path of the old side-chain. The maximum number of bond lengths, angles and torsion angles are taken from the old side-chain that are compatible with the new side-chain. Additional internal co-ordinates to complete the side-chain are taken from the secondary structure dependant rotamer library (McGregor et al., 1987) as described in (Bates et al., 1997). After the replacement of all side-chains and the assignment of a single rotamer for each, this parent rotamer, plus rotamers from a side-chain rotamer library are built at each residue position. Each rotamer feels the average environment due to rotamers of other residues weighted by their respective probabilities. The conformational probability matrix is refined to give resultant probabilities for the rotamers of each residue

(Koehl & Delarue, 1994). Energy parameters taken from (Lee & Subbiah, 1991). The highest probability rotamer was chosen. The selected rotamers were inspected to check that all rotamers look sensible. If not then the side-chain rotamer from just the first stage in the side-chain replacement mechanism is used. Only a few rotamers needed to revert back and these were usually polar or charged residues involved in conserved hydrogen bonding interactions. The reasons for the mean field not always finding the best rotamer is addressed. To remove the small number of steric clashes remaining in the models 100 steps of steepest descents energy minimization (unrestrained) were run, on all models, using the program CHARM (CHARM software version 3.3 (1992) Molecular Simulations Inc. 200 Fifth Avenue, Waltham, MA 02154.). We have submitted unrefined and refined coordinates, before and after energy minimization. The significance of performing unrestrained global minimization, irrespective of the expected modelling errors, is addressed. References Bates, P.A., Jackson, R.M. Sternberg, M.J.E. Model building by comparison: A combination of expert knowledge and computer automation. Proteins, Suppl. 1,59-67, 1997. Gerstein, M., Levitt, M. Using interative dynamic programming to obtain accurate pairwise and multiple alignments of protein structures. Fourth International Conference on Intelligent Systems for Molecular Biology. AAAI Press, Menlo Park, CA, USA, 59-67, 1996. Jones, T.A., Thirup, S. Using known substructures in protein model building and crystallography. EMBO J. 5, 819-823, 1986. Koehl, P., Delarue, M. Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformation entropy. J. Mol. Biol. 239, 249-275, 1994. Koehl, P., Delarue, M. A self consistent mean field approach to simultaneous gap closure and side-chain positioning in homology modelling. Nat. Struc. Biol. 2, 163-170, 1995. Lee C., Subbiah S. Prediction of protein side-chain conformation by packing optimization. J. Mol. Biol. 217, 373-388, 1991. McGregor, M.J., Islam, S.A., Sternberg, M.J.E. Analysis of the relationship between side-chain conformation and secondary structure in globular proteins. J. Mol. Biol. 198, 295-310, 1987. Olivia, B., Bates, P.A., Enrique, Q., Aviles, F.X., Sternberg, M.J.E. An automated classification of the structure of protein loops. J. Mol. Biol. 266,814-830, 1997.

Prediction of protein structure using contacts predicted by neural networks

Ole Lund (1), Jan Gorodkin (1), Claus A Andersen (1), Jakob Bohr (2), Soren Brunak(1). (1) Center for Biological Sequence Analysis, (2) Department of Physics, The Technical University of Denmark, DK-2800 Lyngby, Denmark.

Contacts in proteins were predicted using the SoWhat method (O. Lund, K. Frimand, J. Gorodkin, H. Bohr, J. Bohr, J. Hansen, and S. Brunak. Protein distance constraints predicted by neural networks and probability density functions. Prot. Eng., 10: 1241-1248, 1997). In brief, a neural network were used to predict whether two amino acids were in contact or not. Two residues were defined as being in contact if their C-alpha atoms were closer than the average distance between residues at the given sequence separation. Contacts were also predicted using an updated version of the SoWhat method where the window sizes has been optimized guided by measurements of the information content of the input sequences. These predictions were submitted to casp3 as MODEL 2. Folds were recognized using the FASTA program. If no significant match were found the SwissProt entries of the top scoring matches were inspected manually to identify probable evolutionary relationship. Sequences were aligned using an alignment procedure which disallowed insertions and deletions between two residues in the same secondary structure element (alpha-helix, beta-sheet) in the template, and between residues i < j, where the distance between i-1 and j+1 were greater than 10.5 AA. Three dimensional models were build by homology modeling using the distance constraints predicted by neural networks (the SoWhat method) and folding using the FOLD program (Lund O, Hansen J, Brunak S, Bohr J. Relationship between protein structure and geometrical constraints. Protein Sci 1996 Nov; 5(11):2217-2225). C$^\alpha$

atoms only, were used to construct the cost function for the folding program which uses a simple steepest descent algorithm to minimize the cost function. The above described calculations are available via the CPHmodels server at http://www.cbs.dtu.dk/services/CPHmodels/

Handling interconnected structural changes in comparative modelling of proteins using a statistical scoring function, graph theory, and exhaustive enumeration techniques

Ram Samudrala and Michael Levitt

The interconnected nature of interactions in protein structures, thorough sampling of side chain and main chain conformations, and devising a discriminatory function that can distinguish between correct and incorrect conformations are the major hurdles preventing the construction of accurate homology models. We present an algorithm that uses graph theory to handle the problem of interconnectedness. Sampling of side chain and main chain conformations is accomplished by exhaustively enumerating all possible choices using a discrete state model. The optimal combination of these possibilities is selected using an all-atom scoring function. Following is a brief description of the components of this method: 1. Discriminatory function is an all-atom distance-dependent conditional probability discriminatory function based on a statistical analysis of known protein structure. The negative log of the conditional probability of observing two atoms interact given a particular distance is used as a ``pseudo-energy" to assign weights to nodes and edges. Reference: J Mol Biol 275: 893-914 (1998). 2. Side chains possibilities are generated by selecting the most probable side chain rotamers based on the interactions of a given rotamer with the local main chain (evaluated using the discriminatory function above). Reference: Prot Eng 11: (1998). (in press) 3. Main chain sampling is performed using an exhaustive enumeration technique based on discrete states of phi/psi angles. (unpublished) 4. The graph theoretic approach represents each possible conformation of a residue in an amino acid structure is represented using the notion of a node in a graph. Each node is given a weight based on the degree of the interaction between its side chain atoms and the local main chain atoms. The weight is computed using a all-atom conditional probability discriminatory function. Edges are then drawn between pairs of residues/nodes that are consistent with each other (i.e., clash-free and satisfying geometrical constraints). The edges are also weighted according to the probability of the interaction between atoms in the two residues. Once the entire graph is constructed, all the maximal sets of completely connected nodes (cliques) are found using a clique-finding algorithm. The cliques with the best probabilities represent the optimal combinations of mixing and matching between the various possibilities, taking the respective environments into account. Reference: J Mol Biol 279:287-302 (1998). 5. Clique-finding is accomplishing using the Bron and Kerbosch algorithm. Reference: Communications of the ACM, 16: 575-577 (1973). We test how the above approach works in a comparative-modelling scenario and assess the predictive power of this method by applying it to properly controlled blind tests as part of the third meeting on the Critical Assessment of protein Structure Prediction methods (CASP3). Compared to CASP2, where a similar approach was used, we have dramatically changed the method used to sample main chains (database vs. exhaustive) and have made minor enhancements to the other components of this approach. It remains to be seen how the improvements in methodology correlate with model accuracy.

Protein Secondary Structure Prediction Using Optimised Nearest Neighbor Method

G. P. S. Raghava

One of the challenge in the field of protein structure prediction is to improve the accuracy of secondary structure prediction. Presently, available secondary structure prediction methods, can be classified into i) statistical methods; ii) physiochemical methods; iii) Artificial Intelligence (AI) based methods; vi) multiple sequence alignment method and v) Combinatorial methods. The artificial intelligence (AI) based protein secondary structure prediction methods can be categorize into neural network methods and nearest neighbor methods. Both approach have their advantage and disadvantage. The basic idea of nearest neighbor methods is to use the examples closely related to the test instance to determine the secondary structure of the test instance. The success of prediction of this approach is directly depend upon the closely related known examples corresponding to a test instance. The nearest neighbor method outperform the neural network if their are similar or identical examples corresponding to a test instances but perform poorly in absence of closely related examples. An optimized nearest neighbor method (ONNM) has been developed to predict the secondary structure of protein from its amino acid sequence. In ONNM, the parameters used in nearest neighbor method have been optimized, to improve the accuracy of prediction of standard nearest neighbor method. In past the database of known examples was limited because these examples were generated from limited set of proteins (around 126) whose structure is known. In order to overcome this limitation, author generate a database of known examples from all proteins in PDB (Protein Data Bank), which maximize the number of examples in database. Author also estimates the probability of correct prediction of each amino acid as well as distribution over the three states. Author also uses the standard neural network method to predict the secondary structure, of amino acids as well as calculate the probability of correct prediction of each amino acid. Finally, the secondary structure of amino acids were predicted using either ONNM or neural network method depending upon which give high probability of correct prediction. This method will be suitable for prediction of proteins who has high homology with proteins of known structure, because it uses the nearest neighbor approach and big database of known examples. It will also be suitable for protein which have very less homology with proteins of known structures because it uses the neural network. The ONNM would be very useful in future because the number of structure solved by X-ray crystallography is increasing every day. Thus, in future their will be number of close related or identical examples of known structures corresponding to each test instance which will improve the accuracy secondary structure prediction. The database of known examples used in ONNM can be update easily, so that information form new structure can be added. The method describe in this report uses only single sequence for prediction.

Ab initio folding by using restraints derived from multiple sequence alignments

Angel R. Ortiz, Andrzej Kolinski, Jeffrey Skolnick

Here we described a new method for protein structure prediction based on ab initio folding using Monte Carlo simulations driven by restraints derived from multiple sequence alignments. Both secondary and tertiary restraints are derived from multiple sequence alignments. Secondary restraints obtained from a prediction of secondary structure. Tertiary restraints are obtained on the form of contact map prediction by a combination of multivariare statistical analysis and inverse folding. The predictions are incorporated as harmonic restraint penalty functions in a reduced protein model of two interacting particles per residue (one mimicking the backbone, the other one the side chain). Interaction between the particles is given by a statistical potential derived from the protein data base, suplemented by the restraint function. Conformationa sampling is carried out on a discretized underlying lattice by a Monte Carlo simulated annealing procedure.

3D Protein Folds: Why and How Homologs Can Help to Predict them.

Alexei V. Finkelstein (1), Azat Ya. Badretdinov (1,2), Boris A.Reva (3) Dmitry S. Rykunov (4) (1) Institute of Protein Research, Russian Academy of Sciences, 142292, Pushchino, Moscow Region, Russian Federation (2) Laboratory of Molecular Biophysics, Rockefeller University, Box 270, 1230, York Ave, New York NY 10021-6399, USA (3) The Scripps Research Institute, Dept. of Mol.Biology, TPC5, 10550 N. Torrey Pines Rd, La Jolla, CA 92037, USA. (4) Institute of Theoretical and Experimental Biophysics, Russian Academy of Sciences, 142292, Pushchino, Moscow Region, Russian Federation

Usually one cannot predict the 3D structure a protein directly from its sequence because of errors in the energy parameters: due to them, the calculated energy of the native fold is often above the calculated energies of some other folds. However, using a set of homologs (the proteins having nearly identical 3D structures despite of numerous amino acid mutations in their chains) one can diminish the errors by averaging the fold energies over the homologs. A simple analytical theory shows that this can allow to predict common fold of the homologs - even when this is impossible for separate sequences. The basic analytical estimate is supported by computer experiments with simplified models of protein chains. The experiments simulate the lowest-energy fold prediction using the "corrupted" energy parameters and show that a sufficiently large set of sufficiently remote homologs allows to recognize this fold correctly. The improvement in protein fold recognition by threading of multiple homologs is also discussed.

KITCHEN SINK FOLD PREDICTION METHOD

Jadwiga Bienkowska, Lihua Yu, Bob Rogers, Jim Freeman, Sophia Zarakhovich Scott Mohr and Temple F. Smith

The fold prediction procedure was done following a set of steps. First using PSI-Blast, we checked the available protein databases for the probable homologs. Second, we generated local profile-induced multi-alignments among all probable homologs. The profile identified conserved residues and multi-alignment gaps. The residue aligned to gap suggests probable surface loop position. We checked the literature for any experiments that may identify function- (and/or structure-) determining residues. Additionally we searched the determined protein structures (PDB) for related or similar biochemical functions. Third, we tried to identify any conserved catalytic and/or protein protein binding sites either via consensus sequence patterns. Using information from the above searches, we assigned (when unambiguous information was available) residues to at least one of three categories: 1) loop or turn residue, probable surface, 2) turn-associated Gly as

probable secondary structure terminal residue, 3) functionally constrained position. Fourth, we ran various secondary structure prediction methods including our own Discrete State Models (DSMs) (C.M. Stultz, J.V. White and T.F. Smith, Protein Science (1993) 2:305-314). The DSMs predict secondary structure in conjunction with fold class. These predictions were compared for consistency and used to identify likely threading templates. Fifth, we used our threading procedure (R.H. Lathrop and T.F. Smith JMB (1996) 255,641-665) with an updated scoring function to select the best sequence-to-structure alignment. Here we included the inter-residue pairwise interaction and used a new algorithm to dynamically filter out the unlikely physical interactions (J.R. Bienkowska R.G. Rogers and T.F. Smith in preparation) . We compared alignments predicted by threading with the secondary structure predictions. We checked for the positions of the conserved GLY (which are probable ends of the secondary structure segments). Any clear inconsistencies between the threading alignment and the above predictions were fixed by manual alignment.

A combined method for low-resolution ab initio tertiary structure prediction of small proteins

Yu Xia, Enoch S Huang, Michael Levitt, and Ram Samudrala

We present a combined approach to construct low resolution models of protein structure from sequence information alone [1]. Starting from protein sequence, we uniformly sample protein conformational space by complete enumeration on a simple lattice and have a pool of up to 10 billion structures. We then use a variety of selection and reconstruction techniques to both reduce the size of candidates and push the distribution of candidate structures to more native-like, till a single structure model of the protein sequence is generated. A detailed description follows: First, we exhaustively enumerate all possible compact bounded lattice walks on a tetrahedral lattice to capture the overall protein topology [2, 3]. The maximum walk length in our approach is 50; hence each vertex can represent more than one residue for a protein of size up to 200 residues. To obtain a model, a lattice walk is threaded with the target protein sequence and the score is evaluated and minimized using a residue-residue contact function. The 10,000 best scoring structures and their mirror images are selected. At this stage only protein tertiary topology is captured; there is no secondary structure or side chain information in these structures. In order to build all-atom models, we fit predicted secondary structures using a combination of three secondary prediction methods (PHD, DSC, Predator) to the selected lattice structures using an off-lattice four-state phi/psi model and a sequential build-up algorithm [4]. The all-atom conformations are minimised using ENCAD and scored using a second scoring function that combined scores produced by three different functions: an all-atom distance-dependent conditional probability discriminatory function [5], a hydrophobic compactness function, and a residue-residue contact function (Shell) [1]. Next, the best scoring 50, 100, and 500 conformations are input to a consensus-based distance geometry routine that uses constraints from each of the conformation sets and produces a single structure for each set (total of three) [6]. We again fit secondary structures to the three structures and the resulting structures are minimised and scored. The lowest scoring conformation is our predicted structure for the give protein sequence. We tested a prediction procedure similar to the one described above on a set of 12 small proteins representing different fold classes [1]. Unlike Monte Carlo and molecular dynamics, our approach is largely deterministic and provides a uniform sampling of structure space. The weakest part in our approach is the low-resolution nature of the tetrahedral lattice model: structural features such as sharp turns cannot be accurately represented. In general, our procedure is more effective for alpha and mixed proteins than beta proteins. [1] Samudrala R, Xia Y, Levitt M, Huang ES. Proc. Pac. Symp. Biocomp., 1999 (accepted). [2]

Hinds DA, Levitt M. Proc. Natl. Acad. Sci. USA, 89:2536-2540, 1992. [3] Hinds DA, Levitt M. J. Mol. Biol., 243:668-682, 1994. [4] Park B, Levitt M. J. Mol. Biol., 249:493-507, 1995. [5] Samudrala R, Moult J. J. Mol. Biol., 275:895-916, 1997. [6] Huang ES, Samudrala R, Ponder J. Prot. Sci., 7:1998-2003, 1998.

Improved and extended local structure predictions using the I-sites Library and a fragment insertion Monte Carlo search algorithm.

Chris Bystroff, Kim Simons, David Baker

A combination of database searching and Monte Carlo (MC) simulation was used in an attempt to predict local and supersecondary structure from single sequences. This is similar but much less cpu-intensive than the procedure used by Simons et al (Proteins, in press, see CASP3 abstract) to predict global tertiary structure. Speed was gained by restraining the residues within ﾓ I-sites ﾒ of high confidence (Bystroff & Baker, J.Mol.Biol 281:565-577, 1998). This method updates the I-sites local structure prediction method usind in CASP2, and is available to the public via the Web at http://ganesh.bchem.washington.edu/~bystroff/Isites. Multiple sequence alignments were made from single sequences using psi-blast , and amino-acid profiles were constructed from the alignments using sequence weighting. All segments of the profile, lengths 5 to 19, were scored against all 261 clusters in the I-sites Library to generate a large set of potential Monte Carlo fragment insertion moves that are consistent with the sequence family. The move set was reduced to the highest-confidence 25 fragments, lengths 3 and 9, for each position in the sequence. Positions in the sequence where the confidence of a fragment exceeded 0.60 were identified, and backbone angle restraints were used for those positions during the MC simulation. (The restraint consisted of simply disallowing any angle change greater than 60 degrees from the restrained value). For longer sequences, the protein was divided up into overlapping fragments . The length of the fragments was chosen so that each fragment contained approximately 50 unrestrained residues. In practice, the fragment lengths were 60 to 120 residues. The points of division were always made in positions of low-confidence I-sites predictions. The sequence to be used in the MC simulation was modified from the target sequence by substituting polar residues for any non-conserved non-polar residues. If the combined frequency of polar residues for any non-polar position exceeded 25%, then the most frequent polar residue was substituted into the sequence and all scoring in the MC simulation was based on this modified sequence. Each fragment was subjected to 4000 cycles of simulated annealing MC fragment insertion using a knowledge-based, all backbone atom potential function described previously. This potential function has been shown to identify near-native conformations in a large (1200) set of decoy structures generated by the MC method . Because of time constraints, only 15 structures were generated for each of the overlapping fragments. The lowest scoring 5 structures for each fragment were selected and combined via an exhaustive genetic algorithm. All crossover points between one fragment set and an adjacent, overlapping fragment set were chosen for fusing any pair of structures. The lowest scoring 5 fused structures were kept and were fused with the resulting structures of adjacent fragment sets until the fused structures were full length. The lowest scoring (energy) structure after this procedure was chosen for submission. There was no manual intervention. Secondary structure assignments, used during the MC simulation, were reported along with the tertiary structure, since often the tertiary structure models are not fully compact and strand pairing remains poorly defined in many cases. The method described here was carried out in two ways, each resulting in one ﾓ TS ﾒ submission and its corresponding ﾓ SS ﾒ submission. In one case the backbone angle restraints were generated so as to be consistent with the highest-reliability PHD helix and strand predictions. (Reliability was converted to confidence according to a formula described in the reference above (Bystroff & Baker, JMB), and a single cutoff, 0.6, was used to select residues to restrain). In the second case, PHD predictions were ignored. Because of the small number of simulations, we expect the predictions to be accurate only at the level of local and perhaps supersecondary structure. In test cases, this method correctly located the positions of most chain reversals but failed to accurately predict strand pairing and other non-local interactions.

Ab initio structure prediction using fragment insertion and simulation annealing

Simons, K.T., Bonneau, R.A., Ruczinski, I. & Baker, D.

To generate structures consistent with both the local and the non-local interactions responsible for protein stability, fragments of known structure closest to the sequence and predicted secondary structure of the target sequence for each overlapping segment of length 3 and 9 were assembled into complete, tertiary structures using a Monte Carlo simulated annealing procedure. (Simons, K.T., et al., JMB 268: 209-25, 1997). The fragments used to build the structures are often very similar to the native conformation; one-half of the 9 residue fragments are within 1?dme of the native structure. The scoring function used in the simulated annealing procedure is composed of sequence dependent terms including such features as hydrophobic burial, electrostatics and disulfide bonds and sequence independent terms capturing the geometrical arrangements between a-helices and b-strands and the formation of b-sheets from b-strands (Simons, K.T., et al., Proteins, in press). For each of 21 small, ab initio targets, 1200 structures were constructed, each the result of 100,000 attempted fragment substitutions. In these sets the ~25 structures with the lowest score in the broadest minima (assessed through the number of structural neighbors; Shortle, D., et al., PNAS 95: 11158-62, 1998) were evaluated by visual inspection. Preliminary results on 5 of the targets indicate that the method is useful for ab initio fold prediction; for the first 90 residues of MarA (116 residues total) the best of the 5 submitted models had an rmsd of 6.6? and for the EH2 domain of EPS15 the best submitted model had an rmsd of 6.0?over the entire 95 residue structured region.

Fold recognition at CASP3 using 3D-PSSMs, SIVA and machine learning protein topology rules

Lawrence Kelley, Bob MacCallum, Gidon Moont, Mansoor Saqi, Marcel Turcotte & Michael Sternberg

Fold Recognition for CASP3 using 3D-PSSMs, SIVA and machine learning protein topology rules Group ID = 3873-9906-1225 e-mail (m.sternberg@icrf.icnet.uk) Lawrence Kelley (1), Bob MacCallum (1), Gidon Moont (1), Mansoor Saqi (2), Marcel Turcotte (1) & Michael Sternberg (1) (1) Biomolecular Modelling Laboratory, Imperial Cancer Research Fund Lincoln メ s Inn Fields, London WC2A 3PX, UK (2) Bioinformatics Group, GlaxoWellcome, Stevenage, UK 1 - CASP3 Entries -------------------- Submission were made for 24 entries in the fold recognition section. Predictions were submitted for: T0043, T0044, T0046, T0051,T0053, T0054, T0056, T0059, T0061, T0062, T0063, T0064, T0065, T0066, T0067, T0068, T0074, T0075, T0077, T0078, T0079, T0080, T0081, T0083 2- Outline ---------- The unknown CASP entry is the target and the library of known folds are the templates. The key aspects of the approach are: (0) Initial check for remote homology of target to templates of known structures using PSI-BLAST. (1) A new method in which multiply-aligned sequences of the target is matched against a multiple sequence alignment constructed from the homologues (or from homologues and analogues) in the SCOP library (3D-PSSM, developed at the ICRF). (2) Local hydrophobicity and predicted secondary structure matched for target and template using SIVA (developed by Bob MacCallum & Janet Thornton, UCL, London). (3) Evaluation of above results in terms of literature and function of target. (4) In addition, we used some of the CASP3 targets to test an approach of scanning predicted secondary structures against topological rules for folds derived by an artificial intelligent machine learning approach (PROGOL, Turcotte, Muggleton & Sternberg). 3 - Template library -------------------- The fold (template) library consists of non-redundant SCOP domains with <40% sequence identity per family (called SCOP40). 4 - Secondary structure predictions ---------------------------------- Two versions of secondary structure prediction were used from multiple-sequence alignments of the target based on PHD (Rost & Sander and DSC (King & Sternberg) implemented in JPRED (Barton). 5 - 3D-PSSMs -------------- The key method used was our new 3D-PSSMs. The concept is that structural superposition of remote homologues can be used to generate a multiple sequence alignment that include more information than can be gathered from sequence searches (such as PSI-BLAST) alone. Each master protein in the SCOP40 library is selected (say A0) in turn. The template 3D-PSSM is constructed from structures within the same SCOP superfamily that can be superposed well in 3D. Based on the structural alignment of say protein domains B0 and C0 to A0, multiple sequence alignments generated by PSI-BLAST for A (A0, A1,A2..Aa), for B (B0,B1,B2..Bb) and for C (C0,C1,C2,..Cc.) are thus equivalenced to generate a multiple alignment: (A0,A1,A2,.Aa,B0,B1,B2,..Bb,C0,C1,C2..Cc). The alignment is converted to a position specific scoring matrix using the approach implemented in PSI-BLAST. Similarly protein B0 and C0 of known structure will be used in turn as masters to construct different 3D-PSSMs. The unknown target is then used to construct a multiple alignment and this is searched in turn against each 3D-PSSM in the template SCOP40 library. The search algorithm is a global dynamic programming algorithm (with an option to penalise end gaps) as developed in FOLDFIT (Russell,R.B., Saqi, M.A.S., Bates,P.A., Sayle,R.A. & Sternberg, M.J.E. (1998). Prot Eng 11, 1-9.) Additional optional features are introduced into the approach. i) A search of an alternate 3D-PSSM library which was constructed from structurally superposing proteins in the same SCOP fold with the aim of recognising either homologues or analogues. ii) The predicted secondary structure of the target is matched against the known secondary structure of the master A0 together with sequence matching in the 3D-PSSM using the approach of FOLDFIT. iii) Weight up the 3D-PSSM score for residues forming the buried core of master A0. iv) Weight up the 3D-PSSM score of totally conserved polar and charged residues as they may be potential functional residues. 6 - SIVA -------- Vector-based alignment of per-residue hydrophobicity and DSC predicted secondary structure probabilities for both target and template. This approach could also be used in the absence of known structures for library sequences. Algorithm is developed by Bob MacCallum & Janet Thornton, UCL, London, unpublished) (7) Visual inspection of results. ------------------------------- The SCOP database was searched to evaluate high ranking hits. Alignments were constructed manually. (8) Submissions ---------------- Most submissions were either a sequence alignments to a known fold or a statement that there was no known fold. The exceptions were: i) T0068 when a 3D model was constructed by Dr Paul Bates, ICRF using the comparative modelling method described in his abstract ii) Target T0064,65,66 when coordinates were submitted. T0064 was constructed from monomers of pdb codes 1r69 and 1lmb ; T0065 from 1lmb and the complex T0066 based on the interactions in the dimer coordinates of 1lmb. 9- Protein Topology Rules derived by Machine Learning -- Some of the CASP3 targets were used to explore our prototype approach of scanning predicted secondary structures against protein topology rules derived using an artificial intelligence based machine learning algorithm (PROGOL) (Turcotte, Muggleton & Sternberg). These rules include data on patterns and types of secondary structures including length, loop length and hydrophobicity.

Threading with contact potentials and models of sequence- and structure conservation

Anna Panchenko, Aron Marchler-Bauer, Stephen H. Bryant

Threading database searches and threading alignment models have been calculated with the method described in "Bryant S.H. (1996) Proteins 26:172-185", "Bryant S.H. et al. (1995) Curr Opin Struct Biol. 5:236-44", and "Madej T. et al. (1995), Proteins 23:356-369", and which already had been used in predictions for CASP1 and CASP2 by this team. Several modifications have been introduced, however, which are discussed below: 1) for most of the models submitted for CASP3 a combined scoring scheme was used: The physics-based contact potential described in "Bryant S.H. et al. (1993) Proteins 16:92-112" has been combined with a quantitative description of sequence conservation within protein families. Based on multiple sequence alignments we construct position-specific scoring matrices (PSSMs) so that the score for aligning a residue at a given position depends on the log-odds of seeing this residue type appearing within the non-redundant set of aligned homologs. Contact-based threading scores and the motif matching scores are combined after refering each to the score distributions obtained for random sequences with the same composition and length (shuffled sequences). We have specified in the REMARK section of each submission which scoring scheme has been used for which model. 2) The statistical significance of contact-based threading scores, and in some cases of PSSM-based and combined scores, is estimated, to rank hits in a database search and to obtain overall confidence in the models generated. We calculate a P-value that reflects the probability with which a random sequence (of the given length and composition) has a score equal or higher than the score obtained for the query sequence. Properties of random-sequence score-distributions are obtained by simulation or by interpolation from reference data (only in case of contact-based scores), relying on the observation that the alignment space size is the major variable in determining random-sequence scores. 3) Sequences are aligned to core structures of structural templates, requiring the minimum core elements to be included in the alignment and allowing for n- and c-terminal extensions up to the inclusion of the whole template structure. These core structures are based on the analysis of intra-domain contacts only if the respective domain does not have a sufficient number of diverse structure neighbors as calculated by VAST ("Gibrat J.F. et al. (1996) Curr Opin Struct Biol. 6:377-385"). The minimum extent of core substructures and the constraints on the size of inserted loops are based on models of structural conservation and divergence within homologous protein families otherwise. Homology is assessed by checking for the presence of "Homologous Core Substructures" (HCS), see "Matsuo Y. et al. (1999) Proteins, in press". In general we have submitted more than one model per prediction target. We have assigned the dummy template "NONE" to model 1 in cases where we have not been able to collect evidence for any other model. We have not submitted predictions for a variety of targets; our goal was to detect remote homologs and we have deliberately excluded "trivial" homologs for which a structural template could be identified by sequence comparison alone (e.g. by BLAST) with high confidence. We have submitted models for a few of the clear homologs though, in cases where we thought that the alignment was ambiguous and an interesting problem per se. We also have excluded targets which appeared too small to form globular structures with a buried hydrophobic core, targets which appeared too large for single domains and where we were not able to do reliable domain parsing on the basis of sequence analysis, and targets with biased sequence composition. We have tried to submit models found at the top of the automated database search hit lists in most cases, details are given in the individual submission REMARKs. The strict limitation on model numbers, however, has forced us to manually edit the hitlists in several cases, and models ranked as number 1 do not necessarily reflect the performance of the fold recognition methods employed. Searching for remote homologs at the borderline of the signal to noise ratio may be an intricate subject, no matter what method is used. We do not think that counting how often a correct template has been identified as the single best scoring hit is an appropriate measure of fold recognition specificity and -sensitivity in the interesting range of "medium to hard" fold recognition problems. We have assumed that models other than "model 1" will be included in the automated evaluation and considered in the assessment of predictions.

PrISM: Protein Informatics System for Modeling

An-Suei Yang and Barry Honig

All alignments and 3D structure predictions for the targets in CASP3 were made with the PrISM program. PrISM (Protein Informatics System for Modeling) is a sequence/structure analysis/modeling system that, used in either interactive or automatic mode, produces 3D structures of proteins from their amino acid sequences. PrISM consists of a variety of integrated computational modules and databases, including the facility to carry out structure topology analysis, sequence homology search/alignment and statistics, structure-structure alignment, multiple sequence/structure alignment, sequence/structure profile analysis, fold recognition, comparative model building, sidechain and loop modeling, and model structure assessment. At present, PrISM make use of SWISS-PROT and PDB as data resources. SWISS-PROT is used without further processing. PDB entries are divided into structural domains with PrISM structure topology analysis tools to form structure domain libraries. These libraries serve as alternative structural databases in addition to the uncut version of the PDB entries. PrISM's sequence search and analysis tools, based on either the Smith-Waterman or Needleman-Wunsch dynamic programming algorithm, in conjunction with statistics based on the theory of extreme value distribution, can perform pairwise sequence similarity searches, pairwise or multiple sequence alignments, sequence family clustering, and sequence profile searches over sequence databases. Structure search and analysis tools use an algorithm which is built upon double dynamic programming and rigid-body superimposition methods. This algorithm is capable of performing pairwise structure alignments, multiple structure alignments, structure similarity searches and clustering over structure databases. The purpose of the sequence and structure analysis modules is to identify the most suitable structural template(s) and to predict the best sequence-to-structure alignments, which are then used in the protein structure modeling modules. Structure templates are recognized first by a Smith-Waterman sequence alignment score computed with the BLOSUM 62 substitution matrix and then normalized using the extreme value distribution. If the sequence similarity score between a query sequence and a template is greater than an empirically determined cut-off of 15, the alignment and the template are used to produce a homology model for the query sequence with PrISM structure modeling modules. If sequence alignment fails to relate a query sequence to any structure in the PDB, a fold recognition procedure is applied. This procedure is started by constructing models (backbone plus carbon beta) of the query sequence based on the predicted alignments of the sequence to all possible templates in the PDB using a sequence-to-structure mapping algorithm. The most likely models for the sequence are then decided by a subsequent model ranking procedure based on a structure fitness score. The structure fitness score is a sum of individual residue scores which are calculated using statistically derived parameters. These parameters are designed to evaluate these simplified models based on secondary structure propensities and the number and chemical properties of the contacting neighbors of each residue. PrISM's structure modeling modules build protein structures using one or more templates that are simultaneously aligned to the query sequence. When more than one template is used, an automatic procedure first divides templates into secondary structure segments, and then selects the most suitable segment templates for model building, segment by segment. Mainchains are built by using the template conformation when possible. Insertion-deletion regions, usually loops, are then rebuilt using ab initio methods. Sidechain torsion angles are either taken from the template or predicted based on the mainchain torsion angles with a neural network algorithm. The model building and the alignment procedures can iterate until a reasonable model structure is arrived. Finally, the model structures are refined with limited minimization or simulated annealing. PrISM contains a model assessment module, which is used to assess the quality of a predicted model as the experimental structure becomes available. The assessment procedure is started by carrying out a structure alignment to align the model and the experimental structure. This is followed by RMSD calculations , the evaluation of the predicted alignment on which the model is built, and the evaluation of the predicted mainchain and sidechain torsion angles. These results provide statistical indicators for the quality of the predicted model. Using PrISM, we have predicted one model for each of the 43 CASP3 targets. The modeling strategy varies from one target to another because the protocol that is used depends on the amount and quality of information extracted from the sequence and structure databases. For example, when a sequence has an EF hand sequence motif but shows very little sequence identity to any structure in PDB, the structure of the sequence is predicted by applying PrISM's fold recognition procedure to a subset of structure database that contains only proteins with at least one EF hand motif. When needed, iterative cycles between information extraction and modeling can be valuable in producing better models. Overall, PrISM provides a flexible computational environment which can be adapted for a large number of different modeling challenges.

Fold recognition from secondary structure predictions

Geetha Vasudevan, Philip G. McQueen, Valentina Di Francesco, Jean Garnier and Peter Munson

The ab initio predictions were limited to the secondary structures using three methods (model 1 to 3) ordered according to the reported accuracy by their authors. SIMPA and QL used homologous sequences to the target sequence when available but GOR (version IV) was used without homologous sequences. SIMPA is a nearest neighbor approach using a similarity matrix, an optimized similarity threshold, a window of 13 to 17 residues, and a database of observed secondary structures. This method is based on the assumption that similar peptide regions have the same secondary structure with a score related to their similarity. (Levin J. et al. FEBS, 205, (1986) 303-308; Levin J. and Garnier J. Biochim. Biophys. Acta, (1988) 955, 283-295.) The updated version used, Simpa96, has an extended database of 324 proteins with Blosum 62 as similarity matrix. Its crossvalidated accuracy ,Q3 , was 67.7% for a single sequence and 72.8% when using multiple alignments of homologues sequences (Levin J. Protein Eng. (1997),10, 771-776). The predicted secondary structures are regularized to a minimum of 4 and 2 residues for helix and strand, respectively with a program developed by Zimmermann K., Protein Eng. (1994), 7, 1197-1202. A confidence scale from 0 to 9 for the prediction at each amino acid position is provided. The Quadratic Logistic (QL) secondary structure prediction method incorporates all single and pairwise residue effects within a 17 amino acid residue stretch upon the predicted secondary structural state of the center residue. Helix and strand periodicity are explicitly considered in this approach. A cross-validated accuracy (Q3) of 62.4% on single sequences and up to 68.5% with homologous sequences is achieved (Di Francesco, Garnier and Munson, Protein Science, vol. 5, pp 106-113, 1996). The GOR, version IV, is an improved version of the GOR method based on information theory with directional information included (Garnier J. et al, J. Mol. Biol., (1978) vol. 120, 97-120). Version IV has an extended database of 267 proteins and counts the frequency of all possible amino acid pairs in a window of 17 residues, its crossvalidated accuracy was Q3 = 64.4% without homologue sequence (Garnier J. et al., Methods in Enzymology, (1996),vol. 266, pp. 540 - 553). The predictions are regularized to 4 residues in helix and 2 in strand and at each amino acid position the estimated probability for each of the three states is printed. Fold recognition was performed with the program FORESST (Di Francesco V. et al., J. Mol. Biol., (1997), 267, 446-463) in which a sequence of predicted secondary structure states of the target protein is aligned to hidden Markov models (HMM) of protein structural families. Secondary structure prediction of the target protein was obtained with the algorithm SIMPA, including homologous sequences when available. Searches were performed against a library of 344 HMMs developed at The Institute for Genomic Research (http://www.tigr.org). These models were trained with sequences of observed structure states of proteins having similar three-dimensional fold, selected from the homology superfamily level of the CATH database (http://www.biochem.ucl.ac.uk/bsm/cath/CATH.html). Z scores were computed for negative log-odds scores assigned by each HMM to the target sequence based on the protein sequences of a control database. The control database consists of 349 sequences of secondary structures predicted with the SIMPA algorithm, for unrelated proteins. Fold prediction of a target sequence was obtained by selecting the structural family described by the HMM for which the following two conditions are satisfied: (1) The Z score for the target sequence should be greater than 2; (2) the model length should differ by less than 20% from the target sequence length. If more than one structural model satisfied these two conditions, the structural family with an activity compatible with that of the target sequence was chosen. A structural template was selected from the proteins used in training the HMM, which best matched the length, secondary structural pattern and if possible the proper alignment of functional residues, using the alignment provided by the HMM.

PROSPECT: a threading-based protein structure prediction system

Ying Xu, Dong Xu, Oakley H. Crawford, J. Ralph Einstein Frank W. Larimer, Michael A. Unseren, Ge Zhang, and Edward C. Uberbacher

We have recently developed a threading-based protein structure prediction system called PROSPECT (PROtein Structure Prediction and Evaluation Computer Toolkit), and have applied the system in the CASP-3 protein structure prediction contest. The core of the system is an algorithm/program that finds a fold-sequence alignment that is guaranteed to be globally optimal, measured collectively by (1) singleton match fitness, (2) pairwise interaction preference, and (3) alignment gap penalties, and does so in an efficient manner. One fundamental difference between our algorithm and others is that we have discovered and used a new parameter C, the topological complexity of a fold template in our formulation of the threading problem. C characterizes the overall structure of pairwise interactions, and is a monotonic function of the cutoff distance for pairwise interactions. It is known that the globally-optimal threading problem changes from being NP-hard to polynomial-time solvable as this cutoff distance changes from a fold's diameter to zero. Some recent research suggests that considering pairwise interactions only between spatially "close" residues is probably sufficient for accurate fold recognition and accurate fold-sequence alignment. We have demonstrated that when the cutoff distance between beta-carbon atoms of the interacting residues is set between 7A to 15A, the computational complexity of our algorithm stays manageable. Mathematically, our threading algorithm finds a globally-optimal fold-sequence alignment in O(nm+ MnN^{1.5C-1}) time, for a fold of m residues and M core secondary structures, and a query sequence of n residues, where N is the maximum number of possible alignments between an individual core of the fold and the query sequence when its neighboring cores are already aligned. C is typically a small positive integer, and it is less than or equal to 6 for 90% of the 1300+ folds in FSSP when the cutoff is set at 7A. In terms of wall-clock time, our algorithm typically takes about two to three days to thread a sequence of 300 amino acids against a database of about 1000 folds on a DEC workstation. To facilitate easy incorporation of known biological constraints and knowledge during the fold recognition process, we have generalized our threading algorithm so that it finds a globally-optimal alignment under various constraints. Currently, PROSPECT allows the incorporation of the following constraints: (1) (partially) known or predicted secondary structures; (2) specified disulfide bonds between certain residues in the query sequence; (3) specified active sites with possibly involved residue species and their geometrical relationship; and (4) specified distance constraints among certain residues in the query sequence. Our algorithm finds a globally-optimal fold-sequence alignment that satisfies the specified constraints. For CASP-3, we ran PROSPECT on each specified target against a database of about 1000 unique folds. Typically the top ten alignments are visually checked and ranked. Then the top one or two alignments are fed into MODELLER to generate atomic models. PROCHECK and WHATIF are used to give the final ranking of the generated atomic models for each predicted structure.

methods of protein homology model construction for CASP3 targets 47,55 and 69.

Mark J Forster

The procedures utilised for the construction of the submitted CASP3 homology modelling targets can be described as the following series of steps. (a) Database search for reference (parent) proteins. (b) Alignment of selected reference proteins with model sequence. (c) Restraint generation and initial model building phase. (d) Evaluation of initial model structures. (e) Refinement of initial models to produce final model. Each of the individual steps used will now be further described. (a) Database search for reference (parent) proteins. Reference or parent proteins were obtained by searching the PDB using the Washington University version of the BLAST program (i.e. WU-blastp) with a BLOSUM62 comparison matrix as implemented on the server of the European Bioinformtics Institute at http://www2.ebi.ac.uk/blast2/. In the case of target 47 (rat alpha(2U) globulin) a single high identity hit (63%) was found against the structure of the mouse major urinary protein (pdb code 1MUP). Other hits observed for this model sequence were close to or below the 30% identity level, did not structurally superimpose on 1MUP, and hence were not used. For target 55, a galactose recognising lectin, the highest scoring hit was an E-selectin (pdb code 1ESL) which was only 29% identical, while the lection domain of mannose binding protein A (pdb codes 1MSBA etc) was 27% identical. Several other hits (1KJB, 1HLI etc) were theoretical models and were not used. This corresponds to a hard homology problem. For target 69 (bovine recombinant conglutinin) all top hits were mannose binding proteins, the top scoring being pdb code 1HUP at 38% identity. A total of four parents were ultimately used (1HUP, 1RTM_1, 1HTN, 1ESL). Targets 55 and 69 were noted to be 23% identical and 28% similar (using GCG gap). (b) Alignment of reference proteins and model sequence. For T047 a Needleman Wunsch alignment vs 1MUP was used but this was trivial due to high identity and no gaps found. For T055 many potential sets of parents and alignment schemes were and the 3D models scored as descibed below. The alignment used for the submitted model was that of parents 1MSB and 1ESL as taken from B.J.Graves et al. (Nature v367, p532 (1994)). The model sequence was hand aligned the parents using the FSSP structural neighbours of 1ESL as a guide. For T069 the alignment of the four parents was that given by the Dali/FSSP database, the model sequence was then hand aligned against these. (c) Restraint generation and initial model building. The insightII 97.0 Homology module (MSI) and interface to Modeler (A.Sali and co-workers) were used to create the above alignments and build initial models. Typically four or eight models were created for each alignment. (d) Evaluation of models. Both initial and refined models were scored by the atomic non local energy analysis (ANOLEA) of Melo and Feytmans (J. Mol. Biol. vol 277, 1141-1152 (1998) as implemented at the site http://www.fundp.ac.be/pub/ANOLEA.html. (e) Refinement of initial models to produce final models. Typically the best initial model, as judged by the ANOLEA e/kT score, was subjected to an in-vacuo minimisation protocol using the Discover program and CVFF force field. The two models submitted for T055 differ in that the first uses a protocol of 2000 steps of steepest descent followed by conjugate gradient with a fixed backbone, until a low maximum energy gradient was achieved. The second model submitted did not fix the backbone for the CG stage. Procheck type geometrical checking was performed using ProStat within insightII 97.0.

Completion and refinement of 3D homology models with Non-Equilibrium Molecular Dynamics

Jaap A. Flohil Herman J.C. Berendsen

The main aim of the PROMOD collaboration [0] for homology modeling is to improve the quality of homology models built by commonly used programs such as WHAT IF or the SwissModel server. Targets 47, 50 and 58 were predicted by homology modeling using single sequence alignment. The BLAST2P [4] server at the EMBL was used to find templates that show at least 50% sequence identity to the CASP modeling competition targets. After side-chain modeling with WHAT IF [2], using the same parameters parameters as the version that is available as homology modeling server at the EMBL in Heidelberg, artefacts of the modeling process have been detected by visual inspection and by automated procedures based on the structure validation modules of WHAT IF. To remove Van der Waals overlap and to adapt to the GROMACS (GROMOS87) force field, an energy minimization was done with GROMACS [1]. In case of target 58, an insertion of a Glycine had to be made at position 34 of template 1AKZ. A Glycine, with a backbone orientation copied from its neighboring residue backbone, and a database allowed restriction on the backbone oxygen, was chosen for insertion. The Carbon-Nitrogen bond between the 2 residues to be separated was removed and replaced by the Glycine, which was shrunk by reducing all its bond lengths to 1/3 of the regular lengths. The so obtained initial model was then regularised applying all forces of the GROMACS force field at 0 K until the lowest local potential energy of the structure is achieved. A 300 ps molecular dynamics simulation in water was performed initially keeping the positions of the crystal waters. A selected fragment of 2 x 3 residues around the insertion were allowed to move, keeping the heavy atom positions of well modeled residues stationary. Every 1 ps frame in the Molecular Dynamics trajectory was analyzed on context dependent backbone counterparts in the PDB and residue packing quality [3], which analyses directional atomic contacts. Only those residues were allowed to move freely that scored poorly in the WHAT IF validation runs or that were selected by visual inspection. Global verification of the backbone structure for each frame was monitored by the Ramachandron Z-score. From the best scoring frame a few residues were selected, in the case of target 58 at 107 ps. Some of the (partly) free residues showing acceptably formed conformations, hydrogen bonds and salt bridges somewhere else in the trajectory, but not optimal present in the best selected frame, were replaced. Residues which showed no improvement after the initial simulation have been exposed to a simulated annealing run of 200 ps, in a 500K - 200K temperature range. This selection process can be reduced: The second run was validated similarly to the previous run, and the residues considered uncertain after the previous run were analyzed. For these residues the frames were selected in which they scored best and those rotamers were carried over to the initial model to create the second model. This second model was energy minimized and submitted. This final minimization was done with X-PLOR [5], as that is most likely also the program used by the experimentalists. Targets 47 and 50 have been modeled with a basically equivalent, but simplified protocol, e.g. without inter-model exchange of side-chains and without visual inspection annotation. Target 58 has been modeled according to the principles of the PROMOD project under the BIOTECH program of the European Commission, coordinated by G. Vriend of the European Molecular Biology Laboratory in Heidelberg, Germany.

REFERENCES 0. The PROMOD partners are: G. Vriend, J. Weare, M.C. Peitsch, N. Guex, L. Holm, H.J.C. Berendsen, J.A. Flohil, S. Hayward, O. Teleman, K. Tappura 1. H.J.C. Berendsen, D. van der Spoel and R. van Drunen GROMACS: A message-passing parallel molecular dynamics implementation Comp. Phys. Comm. 95 (1995) pp. 43-56 2. WHAT IF G.Vriend, WHAT IF: a molecular modelling and drug design program, J. Mol. Graph. 8, 52--56 (1990). 3. Quality Control G.Vriend and C.Sander, Quality control of protein models: directional atomic contact analysis, J. Appl. Cryst. 26, 47--60 (1993). 4. Basic local alignment search tool Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman (1990). J. Mol. Biol. 215:403-10. 5. Brunger, A.T. (1992). X-PLOR Software Manual, Version 3.0 Yale University, New Haven, CT.

Prediction of protein secondary structure at 77% accuracy based on PSIBLAST derived sequence profiles

David T. Jones

In order to assist in the automated prediction of protein structure for genome analysis, a secondary structure prediction method has been developed which makes use of profiles automatically generated by PSIBLAST [1]. Critical to the success of the method is to make use of a carefully filtered non-redundant data bank of protein sequences, which in this case comprised 276030 sequences. Two feedforward neural networks are used to actually make the predictions. The first network predicts secondary structure, the second network filters the output of the first network, in a similar fashion to that used by PHD [2]. The first network has 17 groups of 21 inputs and 75 hidden units, the second network has 17 groups of 3 inputs and 55 hidden units. The networks have been trained on a set of 1887 proteins with an early-stopped training protocol. 10% of the training data was used to detect convergence. Using cross-validation, and a testing set of 160 unique protein folds, the method achieves a Q3 score of just over 77%. References 1) Altschul_SF, Madden_TL, Schaffer_AA, Zhang_JH, Zhang_Z, Miller_W & Lipman_DJ (1997) Nucl. Acids Res. 25:3389-3402. 2) Rost_B (1996) Methods Enzymology (1996) 266:525-539.

Secondary structure prediction by Inductive Logic Programming

Igor Mozetic Center for Applied Mathematics and Theoretical Physics University of Maribor

Inductive Logic Programming (ILP) is an approach to learning from examples. One specifies learning examples in terms of attributes (continuous or discrete) and the result of learning are rules in human-readable form. We have implemented ALF - an ILP system designed to learn from sequential data. ALF is an extension of the well-known learning program C4.5 [1] which produces rules in the form of decision trees. The main advantage of ALF is its flexibility and the ability to efficiently handle large quantities of data (in the order of 100.000 learning examples and 1000 attributes). Flexibility allows the user to easily extend the set of attributes (e.g., by extending the window size), introduce new attributes (e.g., hydrophobicity, polarity, size of residues), or even use previously learned rules as new attributes. The main difference to the neural network or nearest neighbour learning is that the result of learning are explicit rules. A possible disadvantage is that in the rules, only a relevant, non-redundant subset of attributes is used and therefore redundant information cannot be sufficiently exploited. While ALF is a general purpose system for learning from sequential data, we have applied it to the protein secondary structure prediction problem. We have taken a subset of the PDB as specified by the June 1998 release of PDB-select ([2], 25% sequence identity threshold) for which DSSP [3] assigned secondary structure without errors. The subset comprises 887 sequence unique protein chains with 189.718 residues. For all residues we first obtained the PHD server [4,5] secondary structure assignments with reliability index R (0-9). We formed attributes with values of PHD predictions for H if R >= 3, L if R >= 4, E if R >= 5, and unassigned otherwise. We took the window size of +- 6 residues, and additionally formed new attributes in terms of the distance to the nearest H, L, or E residue. In order to prevent overfitting the data, the reduced-error pruning [1] with confidence level CF = 1% was used for learning the rules. 7-fold cross validation on the set of 887 chains yielded Q3 = 72.1 +- 9.7% and SOV [Zemla & Venclovas] = 67 +- 14%. Incremental learning from the whole dataset was then used to submit predictions for the CASP3 targets. References [1] Quinlan,J.R. (1993) C4.5: Programs for machine learning, Morgan Kaufmann. [2] Hobohm,U. & Sander,C. (1994) Enlarged representative set of protein structures, Protein Science 3, 522. [3] Kabsch,W. & Sander,C. (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen bonded and geometrical features, Biopolymers 22, 2577-2637. [4] Rost,B. & Sander,C. (1993) Prediction of protein secondary structure at better then 70% accuracy, J.Mol.Biol 232, 584-599. [5] Rost,B. & Sander,C. (1994) Combining evolutionary information and neural networks to predict protein secondary structure, Proteins 19, 55-72.

Recurrent Bidirectional Neural Network Secondary Structure Predictor Incorporated With Multiple External Secondary Structure Resources (ANACONDA)

Jaap A. Flohil Lab. of Biophysical Chemistry, University of Groningen, Nijenborgh 4, 9747 AG Groningen, The Netherlands Tom de Hoop, Edward E.E. Frietman Delft University of Technology, Faculty of Applied Physics, Lorentzweg 1, 2628 CJ Delft, The Netherlands

The prediction system ANACONDA is trained on amino residues added with their corresponding secondary structure as an input array, which is matched with the target secondary structure during training. During prediction of each residue, the generated secondary structure is returned into the network by a feedback procedure. Evolutionary information for the sequence input of the network is included by applying a PAM 250 matrix scoring function between the residues in the sliding window shifted along the sequence and a secondary structure segment database. The segment database is subdivided in coil, helix, sheet and hairpin fragment groups. The prediction system performs an interactive data exchange between two neural network systems. Predictions start from the N-terminal side. After completion of the sequence, secondary structure data is transferred to the network system which itself is trained on sequences in reversed order. The reverse predictor continues with the reversed sequence and reversed feedback structure array at the C-terminal side. The completed structure array is added to a structure profile and transferred to the forward predictor and the prediction process continues and repeats until self-consistency is achieved. Different from CASP2, the initial assignment for the secondary structure profile is built up using a weighted alignment of secondary structures combining 3 different structure resources: 1) A statistical predictor based on single sequence nearest neighbor algorithm with output filtering. Estimated Q3 accuracy and weight contribution is 60%. 2) A secondary structure threader in which secondary structure segments of the pdbselect* database are threaded along the target sequence using a scoring function based on PAM 250 matrix. The secondary structure of the central residue in the sliding window provided by the highest scoring secondary structure segment in the fragment group database was added to the secondary structure feedback profile. Estimated average accuracy on maximal scoring segments and average weight contribution to the profile is 65%. 3) Blastp[2] is used to align regions of the target sequence to fragments of a number of template proteins from the PDB database. The PDB coordinates of the fragments are translated into dssp secondary structure fragments. This method is applied only when sequence identities are greater then 25% and the fragment length of the aligned regions is greater then 30 residues. Gaps between aligned regions are left open. For every residue in the target sequence a weight factor is added to the structure profile. Weight estimations depend on degree of local sequence identity and structural overlap of sequence fragments. To compensate for the absence of tertiary contacts, weights in the structure profile of the sliding window are adapted according consistency of the previous step between PDB fragments and the structure profile, and memorized for the next step. The (final) secondary structure prediction assignment is post processed from the structure profile by a rule based decision system. ANACONDA has been been designed as a stand-alone prediction system for sequences showing less then 25% homology with database sequences. Since in CASP3 also highly homologous sequences are available as prediction targets, PDB derived secondary structure fragments have been included in the structural feedback profile to obtain more reliable predictions. *June 1998 release of PDB_Select[1] with 901 chains in the 25% sequence identity list 1. U.Hobohm, M.Scharf, R.Schneider, C.Sander: "Selection of a representative set of structures from the Brookhaven Protein Data Bank" Protein Science 1 (1992),409-417. 2. Basic local alignment search tool Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman (1990). J. Mol. Biol. 215:403-10.

Protocol for Protein Structure Prediction

Dave Timms

Step 1. The target sequence is compared to those in a composite sequence database (SWISSPROT-PIR-TREMBL-PDB) using SSEARCH(1). Step 2. The resulting similar sequences are aligned using PILEUP(2). Step 3. Secondary structure is predicted from the alignment using PHD(3)and DSC(4). If identity with templates of known structure > 35% proceed to step 8 Further sequence analysis using Motifs, Hidden Markov Models and coiled coil predictors etc is used to identify any non globular or membranous character which may make fold recognition unsuitable Step 4. The target sequence is threaded against a fold database using THREADER2(5) Step 5. The top 30 hits are re-evaluated in the context of threading results from 50 randomly shuffled versions of the target sequence. Step 6. If no threading hit exhibits a Zscore > 3.0 then re-thread using predicted secondary structure as a constraint. Step 7. The secondary structures suggested from threading results with Z-scores > 3.0 are compared with the secondary structure predictions. Step 8. If the secondary structures correspond, adjust the alignment with the template in the light of the known structure. a)To minimise the structural alterations at insertions or deletions. b)To minimise the replacement of glycines in 'D-conformations' and c)To minimise the introduction of prolines with inappropriate torsions. d)To juxtapose any disulphide partners. Step 9. Exchange template and target sequences using QUANTA(6). Step 10. Introduce insertions via loop library. Step 11. Undertake hierarchical energy refinement using CHARMM(7): a) Anneal deletions and inserted loops,constraining peptide torsions as trans. b) Relieve side-chain clashes c) Aim for minimal movement from original template structure. Step 12. Analyse structure using PROCHECK(8) and WHATCHECK(9) for unlikely geometries etc. Step 13. Perform corrective refinement. If fold prediction is unsuccessful, DRAGON(10) is used for distance geometry folding using constraints derived from secondary structure, disulphide bonds and putative side chain contacts derived from correlation analysis of the multiple alignments. References 1. Pearson W.R., Methods in Enzymology 1996 266: 227-258. 2. Devereux J. et al, Nucleic Acids Research 1984 12: 387-395 3. Rost B. & Sander C., Proteins 1994 19: 55-72 4. King R.D. & Sternberg M.J.E. 1996 Protein Science 5: 2298-2310 5. Jones D.T. 1997 Current Opinions in Structural Biology 1997 7: 377-387. 6. Molecular Simulations Incorporated 1986 7. Brooks B.R. et al, J. Computational Chemistry 1983 4:187-217 8. Laskowski et al, 1993 J. Applied Crystallography 26: 283-291. 9. Hooft R.W.W. et al, Nature 1996 381: 272 10. Aszodi A. & Taylor W.R. 1997 Computers & Chemistry 21: 13-23.

Ab initio and comparative modeling of CASP3 targets using thermodynamic model of secondary structure and distance geometry refinement.

Andrei L. Lomize, Irina D. Pogozheva and Henry I. Mosberg

Ab initio predictions of 3D structure for T0046, T0052, T0059, and T0061 CASP3 targets (gamma-adaptin ear domain, cyanovirin-N, Sm D3 protein, and HDEA, respectively), and a homology-based 3D model of T0048 (pterin-4-alpha-carbinolamine dehydratase) have been submitted. The modeling included the following steps. (1) Identification of tentative alpha-helices and beta-structure, including prediction of beta-sheet topologies and relative shifts of beta-strands, using an incomplete thermodynamic model of regular secondary structure [1]. The model combines free energy terms defining alpha-helix and beta-sheet stabilities in aqueous solution and terms describing immersion of the secondary structures into a nonpolar droplet created by the rest of the protein to calculate, using the dynamic programming algorithm, the lowest energy partition of the peptide chain into alpha-helices, beta-sheets, and nonregular structures. The different free energy contributions to the stabilities of secondary structures were estimated from mutagenesis data, model peptide studies, experimental transfer energies, accessible surface area for interacting side-chains, and using several adjustable parameters. So far, the model has been developed and tested only for alpha-helical peptides and proteins [1]. Therefore, locations of beta-sheets and their topologies were predicted by maximizing continuous nonpolar surfaces of the possible beta-sheets, without quantitative calculations of beta-sheet stability, and assuming that each protein considered (or its subdomain in cyanovirin-N) can form only one beta-sheet. (2) Construction of an initial 3D model by manual (using QUANTA) docking of the predicted helices and the beta-sheets to match their nonpolar areas, provide "knobs into holes" packing of side-chains, and maximize hydrogen bonding. Since this is feasible only for the simplest 3D folds (for example, several alpha-helices packed against a single beta-sheet, a complex of two independently formed beta-sheets, or a single beta-sheet folded in the beta-barrel), we considered only small CASP3 targets with simple all-beta, alpha+beta, or alpha/beta organization predicted in step (1), and omitted more complicated folds, including all-alpha. (3) Distance geometry refinement with evolving system of angle and distance constraints [2]. Each iteration of the refinement procedure includes (1) modification of hydrogen bonding and side-chain torsion angle constraints to eliminate structural flaws found in the previously calculated models, and (2) calculations with DIANA. The flaws (significant distortions of alpha-helix or beta-sheet geometries, buried polar groups with no hydrogen bonds, exposed hydrophobic groups, holes, etc.) were detected using QUANTA and supplementary software [2]. The refinement allows adjustment of the spatial positions and geometries of alpha-helices and beta-sheets and determination of conformations of side-chains, turns, and loops by satisfying, in a stepwise "trial and error" fashion, several simple criteria, such as polarity matching, maximization of hydrogen bonding, close packing, and spatial proximity of residues that appear and disappear simultaneously in multiple sequence alignments [2]. The N-terminal region (first 50 residues) of T0048 has low homology to its parent protein (1dcp) and includes a long insertion. Therefore, we performed a "partial" ab initio prediction (steps 1 and 2) for this region. For the rest of the protein, the initial model was constructed from the structure of the parent, 1dcp. Then, the initial model was refined by distance geometry (step 3). [1] A.L.Lomize and H.I.Mosberg (1997) Thermodynamic model of secondary structure for alpha-helical peptides and proteins. Biopolymers, v.42, pp.239-269. [2] I.D.Pogozheva, A.L.Lomize and H.I.Mosberg (1997) The transmembrane 7-alpha-bundle of rhodopsin: distance geometry calculations with hydrogen bonding constraints. Biophys. J., v.72, pp.1963-1985.

Ab initio and comparative modeling of CASP3 targets using thermodynamic model of secondary structure and distance geometry refinement.

Andrei L. Lomize, Irina D. Pogozheva and Henry I. Mosberg

Ab initio predictions of 3D structure for T0046, T0052, T0059, and T0061 CASP3 targets (gamma-adaptin ear domain, cyanovirin-N, Sm D3 protein, and HDEA, respectively), and a homology-based 3D model of T0048 (pterin-4-alpha-carbinolamine dehydratase) have been submitted. The modeling included the following steps. (1) Identification of tentative alpha-helices and beta-structure, including prediction of beta-sheet topologies and relative shifts of beta-strands, using an incomplete thermodynamic model of regular secondary structure [1]. The model combines free energy terms defining alpha-helix and beta-sheet stabilities in aqueous solution and terms describing immersion of the secondary structures into a nonpolar droplet created by the rest of the protein to calculate, using the dynamic programming algorithm, the lowest energy partition of the peptide chain into alpha-helices, beta-sheets, and nonregular structures. The different free energy contributions to the stabilities of secondary structures were estimated from mutagenesis data, model peptide studies, experimental transfer energies, accessible surface area for interacting side-chains, and using several adjustable parameters. So far, the model has been developed and tested only for alpha-helical peptides and proteins [1]. Therefore, locations of beta-sheets and their topologies were predicted by maximizing continuous nonpolar surfaces of the possible beta-sheets, without quantitative calculations of beta-sheet stability, and assuming that each protein considered (or its subdomain in cyanovirin-N) can form only one beta-sheet. (2) Construction of an initial 3D model by manual (using QUANTA) docking of the predicted helices and the beta-sheets to match their nonpolar areas, provide "knobs into holes" packing of side-chains, and maximize hydrogen bonding. Since this is feasible only for the simplest 3D folds (for example, several alpha-helices packed against a single beta-sheet, a complex of two independently formed beta-sheets, or a single beta-sheet folded in the beta-barrel), we considered only small CASP3 targets with simple all-beta, alpha+beta, or alpha/beta organization predicted in step (1), and omitted more complicated folds, including all-alpha. (3) Distance geometry refinement with evolving system of angle and distance constraints [2]. Each iteration of the refinement procedure includes (1) modification of hydrogen bonding and side-chain torsion angle constraints to eliminate structural flaws found in the previously calculated models, and (2) calculations with DIANA. The flaws (significant distortions of alpha-helix or beta-sheet geometries, buried polar groups with no hydrogen bonds, exposed hydrophobic groups, holes, etc.) were detected using QUANTA and supplementary software [2]. The refinement allows adjustment of the spatial positions and geometries of alpha-helices and beta-sheets and determination of conformations of side-chains, turns, and loops by satisfying, in a stepwise "trial and error" fashion, several simple criteria, such as polarity matching, maximization of hydrogen bonding, close packing, and spatial proximity of residues that appear and disappear simultaneously in multiple sequence alignments [2]. The N-terminal region (first 50 residues) of T0048 has low homology to its parent protein (1dcp) and includes a long insertion. Therefore, we performed a "partial" ab initio prediction (steps 1 and 2) for this region. For the rest of the protein, the initial model was constructed from the structure of the parent, 1dcp. Then, the initial model was refined by distance geometry (step 3). [1] A.L.Lomize and H.I.Mosberg (1997) Thermodynamic model of secondary structure for alpha-helical peptides and proteins. Biopolymers, v.42, pp.239-269. [2] I.D.Pogozheva, A.L.Lomize and H.I.Mosberg (1997) The transmembrane 7-alpha-bundle of rhodopsin: distance geometry calculations with hydrogen bonding constraints. Biophys. J., v.72, pp.1963-1985.

Combining Protein Secondary Structure Predictions Using Machine Learning

Ross D. King, Mohammed Ouali, Arbra T. Strong, David Page

The prediction method is based on the premise that combining different prediction methods will give an improved prediction method - as long as the different prediction methods have some independence of each other. We have therefore chosen to combine the three best three protein secondary structure methods from CASP2 (PHD, DSC, and NNSSP) to form a prediction method that we hope is better than either single method. To do this we formed multiple sequence alignments using the method used in the DSC server for a learning set of 496 non-homologous proteins (selected by Geoff Barton and James Cuff). We collected predictions for these alignments from DSC v1.2, from the PHD server at Heidelberg, and from NNSSP (clu2nnssp v1.2). We were then faced with the problem of how to combine these predictions in the best possible way. This was a particular problem because it was unclear which of our non-homologous sequences PHD and NNSSP had previously been trained on; i.e. we could not be sure that we had an unbiased test set. To overcome this problem we chose a conservative strategy for combining the predictions based on the assumption that the most reliable predictor was PHD, followed by DSC, and then NNSSP. This strategy was to collect the predictions where PHD and DSC differed in their predictions, and to use the machine learning algorithm C5 to find rules to decide which method was correct. The attributes used in the learning were the probabilities and certainty factors of the methods and a set of other descriptors considered relevant. We learned a total of 646 rules, 290 of which tell us to replace PHD by DSC. The estimated accuracy of our method (based on a held-out set) is 73.5%. The estimated accuracy of our rules themselves (by 10-fold cross-validation) is 90%.

CASP3 & FORESST - a library of hidden Markov models of protein structural families for distant homology recognition using secondary structure information

Valentina Di Francesco (1), Maria Cueto (2) and Delwood Richardson (1) (1) The Institute for Genomic Research (2) Novartis

A method for recognizing the three-dimensional structure of a protein from its amino acid sequence based on a combination of hidden Markov models (HMMs) and secondary structure prediction (Di Francesco et al., (1997) J. Mol. Biol. 267: 446-463) has been used for the CASP3 endeavour. Compared to other fold recognition methods based on HMMs, this approach is different in that only secondary structure information is used when training the HMMs. In fact each HMM is trained from known secondary structure sequences of proteins having similar fold selected from the homology superfamily level of the CATH database (http://www.biochem.ucl.ac.uk/bsm/cath/CATH.html). The current library of HMMs, called FORESST, consists of 344 models of structural families, which cover about half of the currently known types of 'unique' structures. To recognize the fold of a target protein, secondary structure states for its amino acid sequence were first obtained with a variety of prediction algorithms (such as PHD, PREDATOR, GOR-IV, SIMPA, etc) in order to increase the chances of having a good prediction on hand. The target predicted sequences were then aligned to each HMM in the library and the predicted fold was the fold described by the model fitting the predicted sequences the best. Three different measures of fitness were taken into account: (a) The HMM scores, i.e. negative log-odds scores normalized by the query sequence length, are an indication of how a HMM of a particular structural family fits the sequence better of a null model. These scores were ranked against the same scores assigned by each model to the predicted secondary structure sequences of 349 unrelated proteins of known 3D structure in a control database. (b) Z scores were computed for negative log-odds (NLO) scores (not normalized) assigned by each HMM to the target sequences relative to the distribution of the NLO scores for the control database. (c) Jscores are an indication of how well a predicted target sequence matches the consensus secondary structure sequence of the proteins in the training set of the HMM of a particular structural family. In the prediction process various other factors were considered, e.g. the model length should not differ much from the target sequence length; the model should fit well most of the predicted secondary structure sequences of the target; the proteins in the predicted structural family should have a biological role compatible with that of the target sequence; the residues involved in the functional activity of the target should be conserved in the chosen structural family. Among the proteins of known 3D structure used in training the HMM, the structural template for the target was selected based on a good agreement in length and pattern of the secondary structures. Often the alignment between template and target was modified to correct for obvious mismatches of secondary structure elements (such as 2 short helices aligned to a long one) or for more reasonable domain parsing. In certain cases, FORESST's predictions were supplemented with profile-searching and motif-finding algorithms such as MoST, ProfileSearch PSI-BLAST, or HMMs trained with amino acid sequences, in order to search for subtle but significant hits that could suggest potential structural and functional relationships. HMMs of premade sequence alignments were built using the HMMER software package. These HMMs were generalized using Blosum62 with higher weight on the matrix than the defaults or using Dirichilet prior mixtures. Alignments of the protein families were sometimes further edited by hand to emphasize conserved features across the family. For a couple of targets (T0062 and T0052) these methods could suggest distant homologues with known 3D structure, and their folds were used as templates for the target sequences. The FORESST methodology described here has already been tested during the CASP2 experiment and was shown there to be successful in a number of cases (Di Francesco et al., (1997) Proteins, Suppl. 1: 123-128). Participation in CASP3 was motivated by the need to test the methodology in real experimental condition with a more extended library of HMMs than the one used in CASP2. It should also be noted that FORESST was used by another CASP3 prediction team (see abstract by Geetha et al., prediction group 3707-4066-9021) by applying pre-set, not changeable criteria for fold prediction with a minimal amount of human intervention. That utilization of FORESST is supposed to emulate a systematic application of the selected criteria for prediction in large fold recognition experiments, such as those for hypothetical proteins from complete genomes.

Fold recognition using ProFIT -- an update

Manfred J. Sippl, Francisco Domingues, Hannes Floeckner, Walter A. Koppensteiner, Peter Lackner, Andreas Prlic, Christian Weichenberger, Markus Wiederstein

ProFIT, the fold recognition tool of the ProCyon package, searches a representative database of protein chains for a possible template structure for a target sequence of a globular protein. The program aligns the given sequence with all protein chains of a data base. The generation of alignments is implemented by a global dynamic programming algorithm and sequence structure compatibility is measured by knowledge based potentials. Here is a summary of the main components of the program. Knowledge based potentials of mean force are derived from the three dimensional structures of proteins [1][2]. Pair interaction potentials are related to radial distribution functions which are extracted from experimental structures by a statistical analysis of atom-atom distances. Surface potentials are derived by an analysis of neighbouring residues in a similar fashion. Two representative subsets of protein chains were extracted from the PDB: one for deriving knowledge based potentials (approx. 400 entries) and a second serving as a fold data base (approx. 2400 entries). Alignment generation is performed by a Needleman-Wunsch dynamic programming algorithm using a substitution matrix derived by application of the frozen approximation and knowledge based potentials [3]. Hence, for every structure in the fold data base a model is generated. The individual models are ranked by a Z-score which takes into account the energy background obtained from a large number of alternative folds. Z-scores are expressed as the difference in energy of a particular model and the average energy background in units of the standard deviation [4]. Fold recognition results were cross-checked with additional information from secondary structure prediction and remote homology searches. Secondary structure predictions were derived from the EMBL PHD server [5] and the program DSC [6]. Remote homology between a target sequence and template structures were detected by our implementation of the intermediate sequence search method [7] which uses gapped BLAST for sequence database searches. References: [1] Sippl MJ, Ortner M, Jaritz M, Lackner P & Floeckner H. Helmholtz free energy of atom pair interactions in proteins. Folding & Design 1, 289-298 (1996) [2] Sippl MJ. The calculation of conformational ensembles from potentials of mean force. An approach to the knowledge based prediction of local structures in globular proteins. J Mol Biol 213, 859-883 (1990) [3] Sippl MJ. Boltzmann's principle, knowledge based mean fields and protein folding. An approach to the computational determination of protein structures. J Comp Aided Mol Des 7, 473-501 (1993) [4] Sippl MJ. Recognition of errors in three-dimensional structures of proteins. Proteins 17, 355-62 (1993) [5] Rost B & Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 232, 584-99 (1993) [6] King RD & Sternberg MJ. Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. Protein Sci 5, 2298-310 (1996) [7] Park J, Teichmann SA, Hubbard T & Chothia C. Intermediate sequences increase the detection of homology between sequences. J Mol Biol 273, 349-54 (1997)

ToPLign Protein Structure Prediction

Ralf Zimmer, Ralf Thiele, Heinz-Theodor Mevissen, Gudrun Lange^*, Jochen Selbig, Thomas Lengauer GMD-SCAI, Schloss Birlinghoven, D-53754 Sankt Augustin, P.O. Box 1316, Germany Tel.: +49 2241 14 2818} e-mail: Ralf.Zimmer@gmd.de ^*: Protein Crystallography, Hoechst Marion Roussel (HMR) AG, Frankfurt, Germany.

For CASP-3, we have made 15 predictions out of the 25 targets we regard as belonging to the fold recognition category with the prediction algorithms implemented in our software tool ToPLign [4]. In particular, we applied three different methods, sequence alignment and (profile) database search, a dynamic programming based algorithm for contact capacity potentials, and a heuristic divide&conquer algorithm to optimize interaction potentials. All three methods align the target sequences against non-redundant representative sets of structures and the whole PDB. By varying important parameters over a set of reasonable settings, these procedures produce several scoring lists, which are ranked according to different criteria. The corresponding alignments are analyzed and reevaluated with various ToPLign tools to come up with a joint 'consistent' prediction of the most probable model structures. Finally, the alignments/threadings with these model structures are refined by tuning alignment parameters using path, energy and reliability profiles as well as parametric optimization algorithms [7] of ToPLign. For the sequence alignment we use standard sequence alignment algorithms with affine gap cost functions with the usual amino acid substitution matrices to compute global, shift, and local alignments. Class context substitution matrices are substitution matrices for pairs of substrings of amino acids instead of pairs of single amino acids. They are derived from multiple and structural alignments of protein families in order to produce a set of family specific matrices.

We estimate statistically whether the alignment score of the target with a respective protein family using the corresponding matrix is significant as compared to the expected score for known members of that family with that matrix. This procedure helps to determine a possible protein family (consisting mostly of proteins of yet unknown 3D structure) for the target sequence in question and to increase the quality of the alignments in the refinement step. The 123D method is tailored towards a fast selection of reasonable targets out of a representative fold library. It is a dynamic-programming based, efficient method for optimally aligning protein sequences to protein structures according to empirical environmental/profile scoring potentials. These so-called contact capacity potentials (CCP) [1] are essentially one-body terms, i.e. depend on one amino acid partner of interactions only and measure the preference of amino acids to have a certain number of contacts in certain structural environments. The potentials are supposed to be generalized measures of hydrophobicity, which is assumed to be an essential driving force for folding proteins into their native structure. The potentials are derived from an analysis of a non-redundant database of highly resolved structures by converting relative frequencies into pseudo energies using a normalization according to the inverse Boltzmann law. In a previous evaluation [1] it has been shown that the scoring function is discriminative enough to recognize native sequence-structure relationships and to detect structural folds in the absence of significant sequence similarity. The recursive dynamic programming (RDP) method [6] is tailored towards the discrimination between alternative plausible structure models and the optimization of the respective alignments. RDP is a heuristic approach for the approximate solution of the full threading problem for a wide range of scoring functions, exploiting sequence conservation, sequence patterns [2], environmental profiles [1], and empirical pairwise interaction potentials [10] of the type introduced by Sippl [5]. Threading problems of this kind are known to be NP-hard [3]. RDP hierarchically assembles locally optimal solutions into partial sequence-structure alignments by focusing on highly conserved regions with highest priority. Conservation can be in terms of local sequence similarities, characteristic sequence patterns, or contact capacity. Then, RDP recursively exploits also interactions derived from already mapped parts (in contrast to the frozen approximation approach) in order to detect weaker signals. RDP is able to produce good fold recognition and, at the same time, biologically reasonable sequence-structure alignments using a modest amount of computing resources.

Acknowledgments: This research has been funded in part by the BMBF funded project "TargId: Methods for the Identification of Protein Targets in Genome Data" under grant no. 0311615.

References: [1] N. Alexandrov, R. Nussinov, and R. Zimmer: "Fast Protein Fold Recognition via Sequence to Structure Alignment and Contact Capacity Potentials", Pacific Symposium on Biocomputing'96, World Scientific Publishing Co., 1996, 53-72. [2] A. Bairoch and P. Bucher and K. Hofmann: "The PROSITE database, its status in 1995", NAR, 189-196, 1996. [3] R. Lathrop: "The protein threading problem with sequence amino acid interaction preferences is NP-complete", Protein Engineering, Vol 7, No 9, 1059-1068, 1994. [4] H. Mevissen, R. Thiele, R. Zimmer, and T. Lengauer: "Analysis of Protein Alignments - The software environment ToPLign", GMD ToPLign WWW interface: http://cartan.gmd.de/ToPLign.html, 1994-96. [5] M. Sippl: "Calculation of Conformational Ensembles from Potentials of Mean Force: An Approach to the Knowledge-based Prediction of Local Structures in Globular Proteins", JMB, 859-883, 1990. [6] R. Thiele, R. Zimmer, and T. Lengauer: "Recursive Dynamic Programming for Adaptive Sequence and Structure Alignment", ISMB'95, C. Rawlings et al. (Eds.), AAAI Press, 384-392. [7] R. Zimmer and T. Lengauer, "Fast and Numerically Stable Parametric Alignment of Biosequences", First Annual International Conference on Computational Molecular Biology (RECOMB'97), M. Waterman (Ed.), ACM Press, to appear 1997. [8] C.Lemmen, A. Zien, R. Zimmer, and T. Lengauer: "Application of parameter optimization to molecular comparison problems", PSB'99, World Scientific, to appear. [9] R. Thiele, R. Zimmer, and T. Lengauer: "Recursive Dynamic Programming", 1998, submitted. [10] Ralf Zimmer, Marko W"ohler, Ralf Thiele: "New Scoring Schemes for Protein Fold Recognition based on Voronoi Contacts", Bioinformatics, 1998, Vol. 14, Nr. 3, 295--308

Comparative Modeling of Targets T0050 and T0060

Ursula Egner

The aim of the study was to refine those procedures for comparative modeling, which were tested at the CASP2 experiment. The structures of the targets were predicted by comparative modeling refining the models by energy minimization techniques. As different refinement procedures influence the final model structures considerably, the refinement protocols used in CASP2 were further tested. Model structures were generated for two targets T0050 and T0060. The sequence identity between the target and the homologous structures in the PDB are 82% and approx. 34%, respectively. The structure of target T050 was determined by NMR techniques. Target descriptions: T0050 (glutamase mutase component S, species: C. cochlearium, 137 amino acids) One homologous structure is known: PDB entry 1be1, sequence identity to target 82%. This protein is a glutamase mutase as well (B12-binding subunit) from C. tetanomorphum. The experimental structure is a minimized average structure from NMR experiments. No indels were observed in the sequence alignment. T0060 (D-dopachrome tautomerase, species: homo sapiens, 117 amino acids) Three homologous structures were identified in the PDB, entries 1gif, 1mif and 1fim with sequence identities to the target sequence of 36%, 36% and 33%, respectively. The model was built based on the 1mif structure, one insertion with respect to 1mif (macrophage migration inhibitory factor from homo sapiens). was found in the sequence alignment. This insertion was placed by the BESTFIT algorithm of the GCG package within a helix. In the model, the insertion was modeled into the loop preceding the helix. A thorough search of candidate loop structures was performed with the loop building facilities of the INSIGHT software from MSI. A template loop was taken from PDB entry 1arb residues B60-B61 corresponding to residues 68-70 of the target. For all other residues, the coordinates were taken from the parent structure and mutated for the correct amino acid residues of the target. The model was built as a trimer as found in related crystal structures, submitted to CASP3 was a monomeric structure. Methods: The software used for sequence alignment was the current GCG package of the University of Wisconsin. The software used for modeling was Insight97.0 and Discover2.97 from MSI Inc. The cff91 force field was applied, with a force constant on the omega angle of 100. The starting structures were soaked with a 9A layer of water. In the first 500 cycles, all heavy atoms were allowed to move. This was followed by a relaxation of all water molecules within a 5A layer around the protein. The outer layer of water molecules was kept fixed during the energy minimization. Finally, the protein and the water molecules were minimized until the maximum derivative was less than 0.5. Three different protocols were tested in the refinement, differing whether the protein atoms were allowed to move freely during the calculations (Model 1) or not (Model 2 and 3). In Model 2, the Calpha atoms in secondary structural elements were tethered with a force constant of 50 and in Model 3, the Calpha atoms in secondary structural elements were fixed.

CALCULATION OF SPATIAL STRUCTURE OF ENGINEERED PEPTIDE RLZ (TARGET T0084)

S.Galaktionov, G.V. Nikiforovich, G.R. Marshall

The procedure starts with the predictions of secondary structure using 12 methods available on internet. All of them predict strong helicity. The consensus pattern contains two longer helical fragments at positions 5 - 19 and 24 - 36. This location of secondary structure was used for prediction of coordination number vector (Rodionov, Mol. Biol.,26,777,1992). The algorithm of reconstruction of contact matrix included the follows elements: 1. Delineation of the constant part Q of contact matrix (contacts between the neighbors, intrahelical contacts/no contacts) and extension of the fixed contact/no contact areas on the basis of probability matrix which elements are functions of combination of "contact" and "no contact" links in the corresponding pathways of length k (k = 2 - 6) determined by Q. 2. Prediction of 3 largest eigenvalues of contact matrix. 3. Reconstruction of sets of eigenvectors 1 - 3 compatible with the extended constant part of contact matrix. 4. Reconstruction of the set of starting matrices from the corresponding eigensystems. 5. Editing starting matrices using specific "goodness criteria" (Galaktionov, Proc. Int. Conf. Syst. Sci., 5, 326, 1994). Predictions of 3D structures were performed on the basis of 4 resulting matrices. The initial reconstruction of spatial structure employed elements of distance geometry. A refinement procedure was oriented at correction of chosen intraglobular distances and removal of stressed contacts by minimization of the corresponding squared deviations with respect to Ca coordinates. The structures 1 and 3 have been found very similar to structures 4 and 2, respectively (RMS< 2 A). Structure 1 has been refined to all-atom resolution. The Ca-trace was used as a template for systematic search for closest low-energy 3D structures (ECEPP/2 force field). The highest number of low-energy backbone structures has been obtained for fragment 1 - 23 (98 conformers); however only one conformer has been found at the level of whole molecule being distinguishingly stable and reasonably close to the template conformation.

Self-Organizing Models of Protein Structure

Robert W. Harrison

We developed an efficient model building algorithm based on Self-Organizing Neural Networks, or Kohonen Networks. There is a strong isomorphism between a linear Kohonen network and a folded polymer chain. The network converges to an ordered self-avoiding structure which satisfies a priori data about the system. This is accomplished with an efficient randomized algorithm. The modified algorithm which includes distance information which greatly enhances convergence. In test cases the implementation in AMMP could readily find compact self-avoiding structures which satisfied distance data. The implementation was tested on

32 CASP3 targets, including 11 ab initio and19 homology models. The ultimate goal for the development of the Kohonen algorithm is ab initio folding of proteins and other polymers. In the absence of long range interaction data, the algorithm readily generates a compact self-avoiding model. However, the model generally has little resemblance to the correct structure. With good distance data, such as NOE derived distances or NOE derived distances in combination with secondary structure, or limited subsets of observed distances the algorithm readily converges to a correct structure. To fold proteins, a good fold recognition potential will need to be integrated with the algorithm. Three simple potentials were evaluated for CASP3. The first was a sequence-based folding potential, where distances were taken from the most similar protein sequences in the PDB. The target distance was the median, and the error bounds were defined by the minimum and maximum distances. Even if this potential were extremely accurate, it would be unsatisfactory because it is not based on first principles. Two hydrophobicity contact potentials were evaluated. One used just alpha carbons; the other added beta carbons. Limited secondary structure predictions were used, where 1-4 distances were restrained for pairs which had well-defined bounds. Since the distance terms alone could converge to either a right handed or a left handed solution, both the initial structure and its enantiomorph were submitted. The difficult part of homology or similarity modeling is treating regions of low homology including insertions, and around deletions. Therefor the network was evaluated as a model-building algorithm for difficult regions. Tests showed that the combination of the network algorithm with a torsion search generated the best models. Five targets were predicted with both "standard" methods and the Kohonen algorithm in order to test the convergence. A new distance restraint generating program was also tested. This program aligned parent structures by the RMS error on each alignment window and used the aligned structures to generate distance terms. Such an approach should be more robust with low-homology families of structures where the individual sequence alignments are less accurate. The goal is to use a family of structures to define a molecular architecture. A major recurring problem in homology/similarity modeling is errors in sequence alignments. The cryptographic technique, known as the incidence of coincidence, combined with dynamic programming, was tested. As implemented, the algorithm aligns sequences when the underlying residue probability distributions are most similar, without explicitly calculating or modeling the probability distributions.

CASP3 & FORESST - a library of hidden Markov models of protein structural families for distant homology recognition using secondary structure information

Valentina Di Francesco, Maria Cueto and Delwood Richardson

A method for recognizing the three-dimensional structure of a protein from its amino acid sequence based on a combination of hidden Markov models (HMMs) and secondary structure prediction (Di Francesco et al., (1997) J. Mol. Biol. 267: 446-463) has been used for the CASP3 endeavour. Compared to other fold recognition methods based on HMMs, this approach is different in that only secondary structure information is used when training the HMMs. In fact each HMM is trained from known secondary structure sequences of proteins having similar fold selected from the homology superfamily level of the CATH database (http://www.biochem.ucl.ac.uk/bsm/cath/CATH.html). The current library of HMMs, called FORESST, consists of 344 models of structural families, which cover about half of the currently known types of 'unique' structures. To recognize the fold of a target protein, secondary structure states for its amino acid sequence were first obtained with a variety of prediction algorithms (such as PHD, PREDATOR, GOR-IV, SIMPA, etc) in order to increase the chances of having a good prediction on hand. The target predicted sequences were then aligned to each HMM in the library and the predicted fold was the fold described by the model fitting the predicted sequences the best. Three different measures of fitness were taken into account: (a) The HMM scores, i.e. negative log-odds scores normalized by the query sequence length, are an indication of how a HMM of a particular structural family fits the sequence better of a null model. These scores were ranked against the same scores assigned by each model to the predicted secondary structure sequences of 349 unrelated proteins of known 3D structure in a control database. (b) Z scores were computed for negative log-odds (NLO) scores (not normalized) assigned by each HMM to the target sequences relative to the distribution of the NLO scores for the control database. (c) Jscores are an indication of how well a predicted target sequence matches the consensus secondary structure sequence of the proteins in the training set of the HMM of a particular structural family. In the prediction process various other factors were considered, e.g. the model length should not differ much from the target sequence length; the model should fit well most of the predicted secondary structure sequences of the target; the proteins in the predicted structural family should have a biological role compatible with that of the target sequence; the residues involved in the functional activity of the target should be conserved in the chosen structural family. Among the proteins of known 3D structure used in training the HMM, the structural template for the target was selected based on a good agreement in length and pattern of the secondary structures. Often the alignment between template and target was modified to correct for obvious mismatches of secondary structure elements (such as 2 short helices aligned to a long one) or for more reasonable domain parsing. In certain cases, FORESST's predictions were supplemented with profile-searching and motif-finding algorithms such as MoST, ProfileSearch PSI-BLAST, or HMMs trained with amino acid sequences, in order to search for subtle but significant hits that could suggest potential structural and functional relationships. HMMs of premade sequence alignments were built using the HMMER software package. These HMMs were generalized using Blosum62 with higher weight on the matrix than the defaults or using Dirichilet prior mixtures. Alignments of the protein families were sometimes further edited by hand to emphasize conserved features across the family. For a couple of targets (T0062 and T0052) these methods could suggest distant homologues with known 3D structure, and their folds were used as templates for the target sequences. The FORESST methodology described here has already been tested during the CASP2 experiment and was shown there to be successful in a number of cases (Di Francesco et al., (1997) Proteins, Suppl. 1: 123-128). Participation in CASP3 was motivated by the need to test the methodology in real experimental condition with a more extended library of HMMs than the one used in CASP2. It should also be noted that FORESST was used by another CASP3 prediction team (see abstract by Geetha et al., prediction group 3707-4066-9021) by applying pre-set, not changeable criteria for fold prediction with a minimal amount of human intervention. That utilization of FORESST is supposed to emulate a systematic application of the selected criteria for prediction in large fold recognition experiments, such as those for hypothetical proteins from complete genomes.

Fold-recognition using hidden Markov models

Christian Barrett, Melissa Cline, Mark Diekhans, Leslie Grate, Kevin Karplus, David Haussler, and Richard Hughey

Group name: UCSC-compbio (9070-5088-8627) Authors: Christian Barrett, Melissa Cline, Mark Diekhans, Leslie Grate, Kevin Karplus, David Haussler, and Richard Hughey Computer Science and Computer Engineering Departments University of California, Santa Cruz Faculty leader: Kevin Karplus Student leader: Christian Barrett Fold recognition was performed using the Target98 (SAM-T98) method [3] using SAM version 2.1.1 [1], a refinement of the methods developed by this group for CASP2 [2]. This method attempts to find and multiply align a set of homologs to a given sequence, then create an HMM from that multiple alignment. Each of the following general steps are discussed further below. First, an iterative method is used to create a multiple alignment of homologs for a given target sequence. Sequence weights are then calculated for the this multiple alignment. Next, SAM's modelfromalign program is used to build a model from the alignment and the sequence weights. Finally, SAM's hmmscore performs a local, all-paths scoring of the sequences, using a reversed-model normalization feature [3]. The weighting method, detailed in upcoming publications [3,4,7], combines the Henikoffs' scheme [5], Dirichlet mixtures [6], and an entropy method to set the final weights. Alignment generation While a more detailed description of this process will be available in an upcoming publication [7], we provide the basic outline of it here. The initial step uses BLASTP to search NRP to get two sets of homologs: one of very similar sequences that are almost certainly homologs, and one of vaguely similar sequences. The second set should contain most of the homologs that can be found by any sequence- or profile-based method, but also contains a large number of false positives. The method then uses multiple iterations of a selection, training, and alignment procedure. Each iteration involves an initial alignment, a set of search sequences, a threshold value, and a transition regularizer. The first iteration uses a single sequence (or seed alignment) as the initial alignment and the close homologs found by BLASTP as the search set. The threshold is set very strictly, so that only strong matches to the sequence are considered. This iteration uses a transition regularizer that was designed to match the gap costs used by BLASTP. On subsequent iterations the input alignment is the output from the previous iteration, the search set is the larger set of possible homologs found by BLASTP, and the thresholds are gradually loosened. The second through second-from-last iteration use a ``long-match'' transition regularizer, and the final iteration uses a transition regularizer trained on FSSP alignments. CASP3 We carefully examined the top few predictions of the automatic method, often hand-editing the alignments produced. The visual examination of the prediction and hand-editing of the alignment were the only places that structural information were included in the method. The automatic part of the method is available on http://www.cse.ucsc.edu/research/compbio/HMM-apps/ With few exceptions, we made one prediction for every CASP3 target. For some targets, we had no strong predictions, but chose to make a highly speculative one anyway. For each of these, we provided a probability that it is a false positive, based on our previous fold-recognition tests [7]. References [1] Richard Hughey and Anders Krogh, Hidden Markov models for sequence analysis: Extension and analysis of the basic method. CABIOS 12(2): 95-107, 1996. http://www.cse.ucsc.edu/research/compbio/sam.html [2] K. Karplus, K. Sjolander, C. Barrett, M. Cline, D. Haussler, R. Hughey, L. Holm, and C. Sander, Predicting Protein Structure using Hidden Markov Models. Proteins: Structure, Function, and Genetics, Suppl. 1, 134-9, 1997. [3] Kevin Karplus, Christian Barrett, Richard Hughey, Hidden Markov Models for detecting Remote Protein Homologies. To appear in Bioinformatics. [4] J. Park, K. Karplus, C. Barrett, R. Hughey, D. Haussler, T. Hubbard, and C. Chothia, Sequence Comparisons Using Multiple Sequences Detect Twice as Many Remote Homologues As Pairwise Method. http://cyrah.med.harvard.edu/assess_final.html to appear in JMB, 1998. [5] Steven Henikoff and Jorja G. Henikoff, Position-based Sequence Weights. JMB, 243(4), pp 574-578, Nov 1994. [6] K. Sjolander, K. Karplus, M. P. Brown, R. Hughey, A. Krogh, I. S. Mian, and D. Haussler, Dirichlet Mixtures: A Method for Improving Detection of Weak but Significant Protein Sequence Homology. CABIOS 12(4):327-345, 1996. [7] Kevin Karplus, Christian Barrett, Richard Hughey, Hidden Markov Models for detecting Remote Protein Homologies. To appear in Bioinformatics.

A Strategy for High Resolution Comparative Modeling

Michal Vieth,Leszek Rychlewski, Valentine J. Klimkowski

Our participation in the competition was predominately limited to the targets with high sequence identity to existing structures in the Protein Data Bank(1). This is in line with our interest in using the structure of the protein target to assist in scoring new compounds, and virtual libraries of possible drug candidates. We strongly believe that only very high quality, structural information of the target is useful in making predictions of binding affinities. Therefore, the most relevant information lies in the arrangement of the binding-site residues. As a consequence, for our intended purposes, the quality of the structure is not necessarily determined by a global structural comparison, but rather by the comparison of the active-site neighborhood. The goal of our participation is to evaluate our approach to comparative modeling, and compare it to recognized, cutting edge methods currently available. The general strategy used involved the following steps: I) For each target considered, homologous sequences were identified from the PDB by BLAST(2) and FASTA(3) searches. Standard parameters were applied as given in the Biology Workbench(4). II) The determined percent identity and number of gaps were used to select the best FASTA alignments. III) Modeller(5), with partial refinement run from the Quanta97(6) interface, was used to generate the models. Depending on the target being constructed, one or more parent structures were chosen for model building. IV) The resulting model structures were each then evaluated by the Protein Health module in Quanta97; the Luthy-Eisenberg 3D profile scores(7); and then by visual inspection. V) Each model was then energy minimized with CHARMM(8) using the param19/toph19 force field. VI) The structures were then subjected to the iterative refinement procedure described in detail in Vieth et al(9). The key feature of the referenced refinement protocol is the iterative molecular dynamics, simulated annealing in the presence of a water shell while incorporating restraints to preserve secondary structure elements. At the end of each stage of refinement, the resulting structures were averaged, and then energy minimized. The resulting minimized average structure was then used as the starting point for the next iteration of refinement. VII) After three to six applications of the refinement protocol, the final structures were used to determine their Luthy-Eisenberg 3D profile score, and also the number of interior holes. The structures having a similar number of holes were ranked by their Luthy-Eisenberg scores. In three cases (target t0084, target t0056, and second domain of target t0069) there were no homologous structures found in the PDB. Target t0084 showed high probability to form a helical coiled coil structure(10). Further examination of this sequence revealed a 98% sequence identity to a retro sequence of the leucine zipper (rlz) from the GCN4 transcriptional activator(11). The associated structure was subsequently used for model building of the two stranded, helical coiled coil using the refinement protocol described above. The second domain of target t0069 also showed a high probability of forming a coiled coil structure. The similarity of the first domain of t0069 to the PDB entry 1hup, whose second domain forms a three stranded coiled coil, prompted us to model the second domain of t0069 as a three stranded helical coiled coil. The orientation of the two domains was modeled based on the structure of 1hup. It is our expectation that the results of the competition will be of benefit not only our group ✗ s specific interests, and also others trying to determine what approaches are most successful in comparative modeling. References 1)Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M., The Protein Data Bank: a computer-based archival file for macromolecular structure,J. Mol. Biol. 1977, 112, 535-542. 2)Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J., Basic local alignment search tool,J. Mol. Biol. 1990, 215, 403-410. 3)Pearson, W. R.; Lipman, D. J., Improved tools for biological sequence comparison,PNAS 1988, 85, 2444-2448. 4)Subramanian, S. Biology Workbench v.1.5; Subramanian, S., Ed.; University of Illinois: Urbana-Champaign, 1996. 5)Sali, A.; Blundell, T. L., Comparative protein modelling by satisfaction of spatial restraints,J. Mol. Biol. 1993, 234, 779-815. 6)MSI QUANTA; 4.6 ed.; MSI, Ed.; Molecular Simulations Inc.: San Diego, CA, 1997. 7)Luthy, R.; Bowie, J. U.; Eisenberg, D., Assesment of protein models with 3-dimensional profiles,Nature 1992, 356, 83-85. 8)Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M., CHARMM: A program for macromolecular energy, minimization and dynamics calculations,J. Comp. Chem. 1983, 4, 187-217. 9)Vieth, M.; Kolinski, A.; Brooks, C. L. I.; Skolnick, J., Prediction of the folding pathways and structure of the GCN4 leucine zipper,J. Mol. Biol 1994, 237, 361-367. 10)Lupas, A.; Van Dyke, M.; Stock, J., Predicting Coiled Coils from Protein Sequences,Science 1991, 252, 1162-1164. 11)O'Shea, E.; Klemm, J. D.; Kim, P. S.; Alber, T., X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled coil,Science 1991, 254, 539-544.

Suite of threading and profile-based methods for structure prediction

Godzik A, Jaroszewski L, Pawlowski K, Rotkiewicz P, Rychlewski L, Zhang B

Secondary Structure prediction: We have utilized two methods: PHD (Rost & Sander, PNAS, 1993, 90, 7558-7562; Rost & Sander, JMB, 1993, 232, 584-599; and Rost & Sander, Proteins, 1994, 19, 55-72) and the NO NAME method (Rychlewski L, Zhang B and Godzik A, "Searching for the optimal sequence similarity function", Protein Eng, in revision) available on our server (http://cape6.scripps.edu/). Both predictions are combined by adding both reliabilities obtained for prediction of the helix, coil and extended state and choosing the state with the highest sum. The NO NAME method is based on a nearest neighbor approach. For the query sequence and all sequences in the fold database an alignment with homologous proteins is used to generate profiles. The profiles based on vectors of frequencies of 20 amino acids are transformed into profiles based on vectors of 6 features. The transformation procedure was optimized for highest secondary structure prediction accuracy. Every profile is divided into overlapping 12-residue long segments. For every query segment the set of 40 most similar segments from the fold database is determined. The secondary structure distribution in this set of similar segments is used to estimate the preferences of the query sequence. Fold prediction: Two methods: Hybrid Threading (Jaroszewski L, Rychlewski L, Zhang B and Godzik A, "Fold predictions by a hierarchy of sequence, threading and modeling methods" Protein Science 7:1431-1440) and BASIC (Rychlewski L, Zhang B and Godzik A, "Function and fold predictions for Mycoplasma genitalium proteins", Fold Des 1998 3:229-238) are used to find and validate sequence structure compatibility. The Hybrid Threading method utilizes sequence similarity, secondary structure prediction, burial state prediction, and R14 prediction (distance between C-alpha of residue i and i+4) to estimate the sequence-structure fitness between the query protein and a fold in the database. BASIC (available on http://cape6.scripps.edu/) is a sequence-only profile-to-profile alignment method (gaps are allowed). Profiles are built using PSI-BLAST to find homologous sequences for the query sequence as well as for the proteins in the fold database. Homology modeling: MODELLER (Sali A and Blundell TL, J Mol Biol 234, 779-815, 1993) is used based on alignments obtained with BASIC.

Suite of threading and profile-based methods for structure prediction

Godzik A, Jaroszewski L, Pawlowski K, Rotkiewicz P, Rychlewski L, Zhang B

Secondary Structure prediction: We have utilized two methods: PHD (Rost & Sander, PNAS, 1993, 90, 7558-7562; Rost & Sander, JMB, 1993, 232, 584-599; and Rost & Sander, Proteins, 1994, 19, 55-72) and the NO NAME method (Rychlewski L, Zhang B and Godzik A, "Searching for the optimal sequence similarity function", Protein Eng, in revision) available on our server (http://cape6.scripps.edu/). Both predictions are combined by adding both reliabilities obtained for prediction of the helix, coil and extended state and choosing the state with the highest sum. The NO NAME method is based on a nearest neighbor approach. For the query sequence and all sequences in the fold database an alignment with homologous proteins is used to generate profiles. The profiles based on vectors of frequencies of 20 amino acids are transformed into profiles based on vectors of 6 features. The transformation procedure was optimized for highest secondary structure prediction accuracy. Every profile is divided into overlapping 12-residue long segments. For every query segment the set of 40 most similar segments from the fold database is determined. The secondary

structure distribution in this set of similar segments is used to estimate the preferences of the query sequence. Fold prediction: Two methods: Hybrid Threading (Jaroszewski L, Rychlewski L, Zhang B and Godzik A, "Fold predictions by a hierarchy of sequence, threading and modeling methods" Protein Science 7:1431-1440) and BASIC (Rychlewski L, Zhang B and Godzik A, "Function and fold predictions for Mycoplasma genitalium proteins", Fold Des 1998 3:229-238) are used to find and validate sequence structure compatibility. The Hybrid Threading method utilizes sequence similarity, secondary structure prediction, burial state prediction, and R14 prediction (distance between C-alpha of residue i and i+4) to estimate the sequence-structure fitness between the query protein and a fold in the database. BASIC (available on http://cape6.scripps.edu/) is a sequence-only profile-to-profile alignment method (gaps are allowed). Profiles are built using PSI-BLAST to find homologous sequences for the query sequence as well as for the proteins in the fold database. Homology modeling: MODELLER (Sali A and Blundell TL, J Mol Biol 234, 779-815, 1993) is used based on alignments obtained with BASIC.

Suite of threading and profile-based methods for structure prediction

Godzik A, Jaroszewski L, Pawlowski K, Rotkiewicz P, Rychlewski L, Zhang B

Secondary Structure prediction: We have utilized two methods: PHD (Rost & Sander, PNAS, 1993, 90, 7558-7562; Rost & Sander, JMB, 1993, 232, 584-599; and Rost & Sander, Proteins, 1994, 19, 55-72) and the NO NAME method (Rychlewski L, Zhang B and Godzik A, "Searching for the optimal sequence similarity function", Protein Eng, in revision) available on our server (http://cape6.scripps.edu/). Both predictions are combined by adding both reliabilities obtained for prediction of the helix, coil and extended state and choosing the state with the highest sum. The NO NAME method is based on a nearest neighbor approach. For the query sequence and all sequences in the fold database an alignment with homologous proteins is used to generate profiles. The profiles based on vectors of frequencies of 20 amino acids are transformed into profiles based on vectors of 6 features. The transformation procedure was optimized for highest secondary structure prediction accuracy. Every profile is divided into overlapping 12-residue long segments. For every query segment the set of 40 most similar segments from the fold database is determined. The secondary structure distribution in this set of similar segments is used to estimate the preferences of the query sequence. Fold prediction: Two methods: Hybrid Threading (Jaroszewski L, Rychlewski L, Zhang B and Godzik A, "Fold predictions by a hierarchy of sequence, threading and modeling methods" Protein Science 7:1431-1440) and BASIC (Rychlewski L, Zhang B and Godzik A, "Function and fold predictions for Mycoplasma genitalium proteins", Fold Des 1998 3:229-238) are used to find and validate sequence structure compatibility. The Hybrid Threading method utilizes sequence similarity, secondary structure prediction, burial state prediction, and R14 prediction (distance between C-alpha of residue i and i+4) to estimate the sequence-structure fitness between the query protein and a fold in the database. BASIC (available on http://cape6.scripps.edu/) is a sequence-only profile-to-profile alignment method (gaps are allowed). Profiles are built using PSI-BLAST to find homologous sequences for the query sequence as well as for the proteins in the fold database. Homology modeling: MODELLER (Sali A and Blundell TL, J Mol Biol 234, 779-815, 1993) is used based on alignments obtained with BASIC.

Secondary Structure prediction using neural nets and SAM-T98 multiple alignments

Kevin Karplus and Christian Barrett

Group name: UCSC-secondary (1751-3146-3362) Authors: Kevin Karplus and Christian Barrett Computer Engineering and Computer Science Departments University of California, Santa Cruz Neural net prediction We made predictions using only neural nets, even though obvious homologies for some targets could have given much more accurate predictions. All of our secondary structure predictions employed a 4-layer neural net. A window over the previous layer's output was used to generate the input for each layer. The number of units specified the number of different outputs that were available for each layer. Most of our predictions used a neural net with the following structure: layer window units 1 9 6 2 11 9 3 3 8 4 7 3 The three units of the last layer specified the secondary structure, strand(E), helix(H), or coil(C). Network input Our neural net accepts as inputs a multiple alignment generated by the SAM-T98 method, an iterative HMM-building process that is more fully described in our fold recognition abstract and in an upcoming publication [1]. The networks were trained on multiple alignments for each of the proteins that were leaves of the 3-May-98 FSSP tree (except 1cm4A, which has obviously incorrect labeling by DSSP). The training output was defined by DSSP, with H and G (alpha and 3-10 helix) treated as helices, E (beta strand) treated as beta, and everything else as coil. For all of the alignments, sequence weighting and regularization with a 20-component Dirichlet mixture (usually recode3.20comp) were applied to arrive at amino acid probabilities for each column of the multiple alignment. The relative weights of the sequences were set by the Henikoffs' weighting scheme [2], and the absolute weights were set so that the regularized distributions of all columns averaged 1-bit of savings relative to the background distribution. These sequence weights were used to generate weighted amino acid counts for each column, which were regularized to probabilities using the same Dirichlet mixture as above. These resulting probabilities were the input to the neural network, as were the weighted probabilites of deletions or of insertions before columns. References [1] Kevin Karplus, Christian Barrett, Richard Hughey, Hidden Markov Models for detecting Remote Protein Homologies. To appear in Bioinformatics. [2] Steven Henikoff and Jorja G. Henikoff, Position-based Sequence Weights. JMB, 243(4), pp 574-578, Nov 1994.

Bayesian model (profile-like model) for remote homology search

Jun Zhu, Roland Luethy and Charles Lawrence

This method is for fold recognition. The target is compared with PDB25 database (U.Hobohm et al., 1992), which is our fold database. For each structure in PDB25, a model is built in the following way: (1) the sequence (query) is compared with NR database using transitive BLAST (Neuwald et al. 1997). Similar sequences are collected and purged at score 150. (2) Then, these sequences are aligned with query sequence using Bayesian aligner (Zhu et al., 1998) and a Bayesian model (profile-like model)is built (Lawrence et al. 1993; Gribskov et al., 1987; unpublished result by Jun Zhu, Roland Luthy and Charlse E. Lawrence). After we have all the models for structures in PDB25, the target file is compared with the models. The result is sorted according to Bayesian evidence of the two not related (Zhu et al., 1998). If the smallest Bayesian evidence is larger than 0.1, we say the target is new fold. Otherwise, we simply choose the structure with the smallest Bayesian evidence to our target as the prediction fold. If we do find target matching to structure in

PDB25, we generate an alignment from marginal posterior probability, as similar in Zhu et al. (1998).

Beta-sheet Prediction Using Inter-strand Residue Pairs and Refinement with Hopfield Neural Network

Minoru Asogawa

The aim of this research is to predict not extend strands but beta-sheetsheets. We gathered statistics of pairs of three residue sub-sequences inbeta-sheets, calculated propensities for them. When a sequence is given,all possible three residue sub-sequences are examined whether they formbeta-sheets. A shortcoming is that many false predictions are made. Thenature of protein tertiary structure precludes the existence of thesefalse predictions. To exclude false predictions and improve theprediction, we employed a Hopfield Neural Network, in which the naturallimitations on protein tertiary structure and preference of chemicallystable long beta-sheet are expressed in a form of energy functions.

Beta-sheet Prediction Using Inter-strand Residue Pairs and Refinement with Hopfield Neural Network

Minoru Asogawa

The aim of this research is to predict not extend strands but beta-sheetsheets. We gathered statistics of pairs of three residue sub-sequences inbeta-sheets, calculated propensities for them. When a sequence is given,all possible three residue sub-sequences are examined whether they formbeta-sheets. A shortcoming is that many false predictions are made. Thenature of protein tertiary structure precludes the existence of thesefalse predictions. To exclude false predictions and improve theprediction, we employed a Hopfield Neural Network, in which the naturallimitations on protein tertiary structure and preference of chemicallystable long beta-sheet are expressed in a form of energy functions.

A self-consistent field based threading algorithm for recognition of protein structure

B.A.Reva, J.Skolnick and A.V.Finkelstein

tervening positions i, i+2 along a chain; (iii) chiral energies depending on a residue pair in positions i+1,i+2 intervening positions i,i+3 along a chain. For each of the target sequences we use 1150 PDB structures as templates. Five lowest energy folds were chosen for prediction submission. REFERENCES 1. Finkelstein, A.V. & Reva, B.A. (1991). A search for the most stable folds of protein chains. Nature 351, 497-499. 2. Finkelstein, A.V. & Reva, B.A. (1996). Search for the most stable folds of protein chains. I. Application of a self-consistent molecular field theory to a problem of protein three-dimensional structure prediction. Protein Eng. 9, 387-397. 3. Reva, B.A. & Finkelstein, A.V. (1996). Search for the most stable folds of protein chains. II. Computation of stable architectures of (-proteins using a self-consistent molecular field theory. Protein Eng. 9, 399-411. 4. Reva, B.A., Finkelstein, A.V., Rykunov, D.S., Skolnick, J. (1998) Optimization of protein structure on lattices with self-consistent field approach. J.Com.Biol., 5, 531-538. 5. Reva B.A, Finkelstein, A.V., Sanner M.F, Olson A.J. (1997) Residue-residue mean-force potentials for protein structure recognition. Protein Eng. 10, pp.865-876.

A self-consistent field based threading algorithm for recognition of protein structure

B.A.Reva, J.Skolnick and A.V.Finkelstein

The threading algorithm used in predictions is based on a self-consistent field theory [1-4]. Details of the SCF-based free energy optimization are given in [4]. The algorithm computes free energy and finds one of the lowest energy structures of a target sequence on a given template structure. In free energy calculation, we take into account all chain conformations including those, which are produced by "loop transitions" between template positions. A loop is given by coordinates of its ends: each of them includes a target sequence coordinate and a 3D coordinate of a template position. No actual loop conformation is considered. Only surface positions on a template can be be connected by loops. Loop lenghs are restricted from 3 to 10 and only the transitions which satisfied a chain connectivity condition are allowed. Due to "loop-transitions", any fragment of a target chain can choose between a position on a template where it contributes to structure determinative interactions, or a position out of a template structure, where it is penalized by non-specific adjustable linear function of a loop length. Thus, a fragmnet of target chain can avoid an unfavorable structural fragment on a template by choosing a transition over the loop. A template structure is given by Ca atom coordiantes. We use Ca atom based distance dependent phenomenological pairwise energy functions which are described in [5]. Distance dependence is approximated by a piecewise constant function which is defined on equal size 1A intervals. With these energy functions, a chain energy is given by a sum of (i) long-distance interactions between residues separated along a chain by 5 or more residues; (ii) local bending energies depending on a residue occuppying position i+1 intervening positions i, i+2 along a chain; (iii) chiral energies depending on a residue pair in positions i+1,i+2 intervening positions i,i+3 along a chain. For each of the target sequences we use 1150 PDB structures as templates. Five lowest energy folds were chosen for prediction submission. REFERENCES 1. Finkelstein, A.V. & Reva, B.A. (1991). A search for the most stable folds of protein chains. Nature 351, 497-499. 2. Finkelstein, A.V. & Reva, B.A. (1996). Search for the most stable folds of protein chains. I. Application of a self-consistent molecular field theory to a problem of protein three-dimensional structure prediction. Protein Eng. 9, 387-397. 3. Reva, B.A. & Finkelstein, A.V. (1996). Search for the most stable folds of protein chains. II. Computation of stable architectures of beta-proteins using a self-consistent molecular field theory. Protein Eng. 9, 399-411. 4. Reva, B.A., Finkelstein, A.V., Rykunov, D.S., Skolnick, J. (1998) Optimization of protein structure on lattices with self-consistent field approach. J.Com.Biol., 5, 531-538. 5. Reva B.A, Finkelstein, A.V., Sanner M.F, Olson A.J. (1997) Residue-residue mean-force potentials for protein structure recognition. Protein Eng. 10, pp.865-876.

Jpred - An automatic consensus secondary structure prediction server

J. A. Cuff, M. E. Clamp and G. J. Barton.

Summary Jpred is an interactive protein secondary structure prediction Internet server. The server allows a single sequence or multiple sequence alignment to be submitted, and returns predictions from six secondary structure prediction algorithms that exploit evolutionary information from multiple sequences. Method The target is searched with BLAST against the OWL v29.4 database, which contains 198,742 entries. The BLAST output (down to a p-value of 0.001) is then screened by SCANPS, an implementation of the Smith Waterman dynamic programming algorithm, with length dependent statistics. Sequences are rejected if their SCANPS probability score is higher than 1x10^-4. Sequences are also rejected if they do not fit a length cutoff of 1.5. For example, if the query sequence is 90 residues long, the sequence length would have to range between 60 and 135 residues to be included. If sequences exceed the length criterion, they are truncated by removing end residues until the length of the sequence satisfies the cut off value. Sequences falling short of the lower length limit are discarded. The value of 1.5 for the length cutoff was reached by visual inspection of a number of multiple sequence alignments, produced with different cut-off values. The method removes both ridiculously long, short and unrelated sequences. However it does allow sequences that are longer than the query, and are related, to be included after truncation. The sequence similar proteins selected by this method, are then aligned by CLUSTALW, with default parameters. The multiple sequence alignments are modified so that they do not contain gaps in the first or 'query' sequence. A slightly different method is used for PHD, whereby only gaps at the end of the target sequence are removed. Without this modification, the conversion of MSF to HSSP file format fails, as a correct insertion table is not constructed. Six different secondary structure prediction methods are then run on the alignment, PHD, DSC, PREDATOR, NNSSP, MULPRED and ZPRED. These methods were chosen as representatives of current state-of-the-art secondary structure predictions methods that exploit the evolutionary information from multiple sequences. Each derives its prediction using a different heuristic, based upon nearest neighbours (NNSSP), jury decision neural networks (PHD), linear discrimination (DSC), consensus single sequence method (MULPRED), hydrogen bonding propensities (PREDATOR), or conservation number weighted prediction (ZPRED). A consensus prediction is calculated by examining the prediction for the four most accurate methods, DSC, PHD, PREDATOR and NNSSP. At each position the prediction is recorded and most popular state is chosen. For example if a residue had the following predictions: NNSSP = Helix, PREDATOR = Helix, DSC = Strand, PHD = Helix the consensus prediction would be Helix. If there was no consensus for a particular residue, the result from the PHD method was used. The Jpred server implements all of the automatic alignment and prediction features described, along with a number of visualisation tools, including a JAVA multiple sequence alignment viewer and editor (JALVIEW). With JALVIEW one can modify the alignment for re-prediction, or interactively change the colouring within the alignment to highlight functionally important residues or conserved features, to aid prediction. For the final prediction the multiple sequence alignment and algorithm predictions were carefully examined by eye, using all of the online visualisation techniques. Simple rules of protein structure were applied to the predictions, such as forcing predicted helices to be at least three residues long. Regions where predictions were ambiguous, and where the alignment contained insertions and deletions were re-considered as indicators of random coil state. Turn predictions from MULPRED were also examined, and used to locate regions where overly long secondary structure elements could be split in to shorter segments. Reliability information from the different predictions was also used to resolve conflicting predictions. Once these filters had been applied, a single line prediction format was generated for CASP3 submission. All confidence values were set at 1.0. http://barton.ebi.ac.uk/servers/jpred.html

---

Prediction of static local backbone deviations in homology models.

Tim Cardozo, Serge Batalov and Ruben Abagyan

We have recently developed an algorithm which assigns a local backbone conservation score to each residue in a homology model. The score, which represents the potential of that residue to deviate in space in its backbone atoms between homologous 3D structures, combines local sequence and local flexibility potential to correctly predict the locations of deviating backbone segments in a homology model 100% better than local sequence or B-factor alone in a benchmark set. The algorithm is here applied tothe CASP3 homology modeling targets and assessed both for its direct accuracy in predicting the precise locations of ungapped loops and gapped loop boundaries and for the improvement knowledge of such boundaries incurs in loop prediction by global optimization methods.

Structure prediction for CASP3 targets

David F. Burke, Nuria Campillo, Charlotte Deane, Mercedes Martin-Martinez, Kenji Mizuguchi, Franck Molina, Hampapathu A. Nagarajaram, Jeff Perry, B.V.B. Reddy, Robert E. Steward, Mark Williams, Joaquim Mendes, Claudio Soares, Maria Armenia Carrondo, Tom L. Blundell

Our predictions for the targets in the Third Meeting on the Critical Assessment of Techniques for Protein Structure Prediction (CASP3) were made using a combination of techniques developed both inside and outside the group. An initial sequence search was carried out on the latest release of the SWISS-PROT and PDB databases using the NCBI WWW PSI-BLAST server[Altschul SF et al. (1997) Nucleic Acids Res. 25:3389-3402]. Secondary structure prediction was performed using the programs DSC[King RD & Sternberg MJE (1996) Prot.Sci. 5(11):298-310], Predator[Frishman D and Argos P (1996) Prot.Eng. 9:133-

142], Sapiens[Wako H & Blundell TL (1994) J.Mol.Biol. 238:693-708], Spetor[Zhu ZY & Blundell TL (1996) J.Mol.Biol. 260:261-276] and the ProteinPredict(PHDsec) server[Rost B & Sander C, (1993) J.Mol.Biol. 232:584-599]. Fold recognition was performed using the UCLA-DOE server[Fischer, D & Eisenberg, D(1996) Protein Sci. 5:947-955] and the program Qslave[Johnson MS et al. (1993) J.Mol.Biol. 231:735-752]. The list of possible structural templates and sequence homologues was compared with suggestions discussed in the literature and the secondary structure predictions. The structure-based alignments of the family and superfamily of the templates were extracted from the HOMSTRAD [Mizuguchi K, et al. (1998) Protein Sci. in press] and CAMPASS [Sowdhamini R et al. (1998) Structure 6(9):1087-94] databases respectively. The target sequence and its close homologues were aligned with the template family using the profile alignment option of CLUSTALW [Thompson JD et al. (1994) Nucleic Acids Res. 22:4673-4680]. This alignment was formatted using the JoY program [Mizuguchi K et al.(1998) Bioinformatics 14:617-623] and examined to give a clearer indication of conserved structural features and conserved amino acid patterns. The alignment was then optimised by hand. The alignment of the templates to the target sequence was used as the basis to the input of the programs MODELLER [Sali A & Blundell TL (1993) J.Mol.Biol. 234:779-815; Sali A et al. (1995) Proteins 23(3):318-326] and COMPOSER [Srinivasan BN & Blundell TL (1993) Protein Eng. 6:501-512]. With MODELLER, several models were generated and the model with the fewest restraint violations was used for further evaluation. Loops were also predicted by searching the SLoop database [Donate L et al. (1996) Protein Sci. 2600-2616; Rufino S et al. (1997) J.Mol.Biol. 267:352-367] and were compared with those built using COMPOSER and MODELLER. Side-chains were placed using the backbone-dependent rotamer side-chain method of Dunbrack as well as an improved version [Mendes et al., submitted] of the Self Consistent Mean Field Theory Method [Koehl P and Delarue M, (1994) J.Mol.Biol.239(2):249-75]. In the latter, interaction energies were calculated with a novel method based on a Flexible Rotamer Model [Mendes et al., in preparation]. For targets with a large number of possible templates, alternative models were produced using only a selection of the templates. The models were validated manually and by using the programs PROCHECK [Laskowski RA et al. (1993) J.Appl.Cryst., 26:283-291],HARMONY [Topham CM et al. (1993) J.Mol.Biol. 229(1):194-220], ProsaII, and Verify3D[Bowie JU et al. (1991) Science 253:164-170; Luthy R et al. (1992) Nature 356:83-85]. New structural alignments were produced using the program COMPARER [Sali A & Blundell TL. (1990) J.Mol.Biol. 212:203-428] and these were used to refine the models.

A new threading algorithm for proteins

Oakley H. Crawford, Ying Xu, Dong Xu, and Edward C. Uberbacher

Threading is a useful technique for predicting the fold of a new protein whose sequence shows no recognized homology with proteins of known structure. The sequence of an unknown protein is aligned to a 3D structure in a way that minimizes a free-energy-like score. This procedure is repeated for all the templates in a fold library. Those templates (if any) giving low enough values of normalized scores are considered to represent likely folds. The threading algorithm has the task of finding the optimal (i.e., lowest score) alignment of a given sequence-structure pair, given a scoring function. The performance of threading algorithms has been a shortcoming in fold-recognition methods. This work assumes a conventional threading model, in which the sequence of an unknown is aligned with the core elements of a template. The scoring function contains singleton and pair terms and a penalty term for indels, which are allowed only in loops. In this case, alignment is in general an NP- complete problem [1]. Thus, any method that always find the global minimum, even in the worst case, requires computational effort that increases exponentially with the number of core elements. Such a method was recently developed by Lathrop and Smith [2]. Xu, et al. [3] have presented an exact threading algorithm that runs in polynomial time when the interactions in the score function are pairwise with a fixed distance cutoff. A new threading algorithm [4] is described here, which ∃ while not guaranteed always to find the global minimum ∃ offers an attractive combination of speed and reliability. The following stochastic search procedure is used. An approximate scoring function consisting of (effective) singleton terms and gap penalties is constructed by replacing the pairwise terms in the score by samples drawn at random from appropriate distributions. A trial alignment is then generated by simple dynamic programming, and its score is subsequently calculated from the original scoring function. The procedure is repeated a predetermined number of times, always drawing new values for the random variables. A limited amount of iteration is employed, to take advantage of information gained as to the likely alignment, but in a way that avoids getting stuck for a long time in a local minimum. The final result is the lowest-score alignment encountered. Comparisons are made with the frozen and iterated frozen approximations [5], which are very fast, and with the exact method of Xu, et al. The new algorithm is found to be considerably more powerful than those approximate methods, and faster than the exact one. ララララララ 1. Lathrop,R.H. (1994). Protein Eng. 7, 1059-1068. 2. Lathrop,R.H. and Smith,T.F. (1996). J. Mol. Biol., 255, 641-665. 3. Xu,Y. and Uberbacher,E.C. (1996). CABIOS 12, 511-517; Xu,Y., Xu,D. and Uberbacher,E.C. (1998). J. Comput. Biol. 5, 597-614. 4. Crawford,O.H. (1998). Bioinformatics, in press. 5. Godzik,A., Kolinski,A., and Skolnik,J. (1992). J. Mol. Biol. 227, 227-238.

Using the CE Structure Neighbors Database to Analyse and Predict Protein Structure Similarity by Sequence

I.N. Shindyalov and P.E. Bourne

The Protein Structure Neighbors Database has been produced from an all-against-all comparison of protein structures in the PDB using the Combinatorial Extension (CE) algorithm [1]. No prefiltering based on sequence similarity has been used. Filtering is performed on structure similarity using a 2A RMSD threshold with 2/3 of residues aligned between two structures. Similarities found were quite different (~60%) than those found by Dali/FSSP [2]. A representative set of structural similarities has been used to calculate an amino-acid substitution matrix. The substitution matrix is similar to the Dayhoff matrix, but with stronger emphasis on hydrophobic interactions. Local dynamic programming alignment of the query sequence with a representative set of unique sequences from the PDB has been performed to reveal compatible structure models. Scores have been normalized to the average score obtained for every representative unique sequence when compared to the sequences from a small set of random folds. The final step was manual selection of the model from the list of top scores. [1] Shindyalov I.N. & Bourne P.E. (1998) Prot. Eng. 11, 739. [2] Holm L. & Sander C. (1993) J. Mol. Biol. 233, 123.

Consensus Approach to Fold Assignment

Robert M. Weiss, Danny W. Rice, Parag Mallick, and David Eisenberg

Some targets required adaptations to our approach, however, most assignments were based on the consensus of three methods: 1. Sequence-derived predictions (Daniel Fischer and David Eisenberg, Protein Science 5:947-955, 1996). This method is based on sequence matching, using a sequence comparison matrix optimized for structural comparisons, but also takes into account the match between observed secondary structure (of the prospective fold) and predicted secondary structure of the target, and uses sequences homologous to the target to aid in the alignment. 2. The modified H3P2 method (Danny W. Rice and David Eisenberg, JMB 267, 1026-1038, 1997), which uses a 3D-1D substitution matrix derived from a set of pairs of homologous structures of low sequence identity, with positions classified by secondary structure type and burial class. The variant of the method used in CASP3 differs from the published one in that the 20 amino acids are treated individually, not divided into seven classes, and a more extensive library of structure pairs was used. 3. Directional profiles, a method currently under development, which (as used in CASP3) looks for patterns in the distribution of sums of atomic solvation parameters within a radius of 14 Angstroms from C-alpha carbons in each of four tetrahedrally arranged directions based on the local path of the main chain. Of these methods, only the first has been calibrated so as to allow a high degree of confidence (in some cases) to be associated with a given prediction. In the process of making a decision, therefore, if the sequence-derived prediction method gave an authoritative prediction, we generally accepted it. Otherwise, we first looked for agreement among the approximately 20 top-ranked predictions between at least two of the three methods. The H3P2 method gives predictions for individual sequences, but also reports them according to their SCOP class. In the interpretation of its results, a higher degree of confidence might be afforded to several top-ranking predictions in the same SCOP class than to a single isolated prediction, even one with higher rank. When information was given or known about a target sequence's function or associated prosthetic groups we typically gave extra consideration to prospective folds that shared those properties. Occasionally, ad hoc tricks were attempted. For example, one fairly long sequence which failed of any prediction was arbitrarily cut in two, and each half considered separately. (When domain boundaries were given by the organizers, the same procedure was applied, less arbitrarily). For difficult targets, we would sometimes also consider the results of Blast searches and use dynamic programming sequence-based methods. One disadvantage of using this farrago of methods is that we probably assigned folds to sequences that should have been assigned to "no known fold". Although we knew that there were probably more sequences in this group than we were predicting, it seemed that, with so many methods available, for each sequence we could come up with some excuse for assigning it to a known fold.

New Methods for Accurate Secondary Structure Prediction, and a Local Pseudopotential for Threading

John-Marc Chandonia, Martin Karplus, and Fred Cohen

Ab initio secondary structure prediction on the CASP3 targets was done using a recently developed neural network based method (Chandonia & Karplus, Proteins, in press). In this paper, a primary and a secondary neural network are applied to secondary structure and structural class prediction for a database of 681 nonhomologous protein chains. A new method of decoding the outputs of the secondary structure prediction network is used to produce an estimate of the probability of finding each type of secondary structure at every position in the sequence. In addition to providing a reliable estimate of the accuracy of the predictions, this method gives a more accurate Q3 (74.6%) than the cutoff method which is commonly used. Use of these predictions in jury methods improves the Q3 to 74.8%, the best available at present. On a database of 126 proteins commonly used for comparison of prediction methods, the jury predictions are 76.6% accurate. An estimate of the overall Q3 for a given sequence is made by averaging the estimated accuracy of the prediction over all residues in the sequence. As an example, analysis is applied to the target b-cryptogein, which was a difficult target for ab initio predictions in the CASP2 study; it shows that the prediction made with the present method (62% of residues correct) is close to the expected accuracy (66%) for this protein. The larger database and use of a new network training protocol also improve structural class prediction accuracy to 86%, relative to 80% obtained previously. Secondary structure content is predicted with accuracy comparable to spectroscopic methods such as vibrational or electronic circular dichroism and Fourier transform infrared spectroscopy. Secondary structure and structural class predictions of CASP3 targets were performed using the Java software available on our web server (http://www.cmpharm.ucsf.edu/~jmc/pred2ary/). Accuracy of the CASP3 target predictions is unknown at this time, but expected to be lower than on the above databases due to lack of multiple homologous sequences for several targets. Predictions done on the CASP3 targets will be discussed at the meeting, and the Java software will be demonstrated. Secondary structure predictions are used to derive a potential energy function representing the conformational preferences of sequentially local regions of a protein backbone. Predictions for each residue in a sequence are used to construct expected distributions of backbone dihedral angles. These are converted into a potential function using the quasichemical approximation. The potential is used in combination with non-local potentials to identify known folds structurally similar (but not identical) to the native structure, and produce global alignments of sequences on these target structures. Preliminary results on test sequences used in previous studies are encouraging. Predictions of sequence alignment for several CASP3 targets, and for larger datasets, will be discussed at the meeting.

Structure prediction using threading, HMMs, and phylogenetic analysis

Kimmen Sjolander and Paul D. Thomas

First of all, we apologize for the length of this abstract. We describe here two methods and their combination. Both of us made predictions for CASP2, independently of each other, and using very different methods. PT developed a threading algorithm while at SmithKline Beecham. KS was on the original team to develop HMMs at UCSC. Since CASP2, we have made substantial changes to each algorithm. The threading algorithm now makes use of predicted secondary structure. The HMMs take subfamily-specific conservation patterns into account. For our CASP3 predictions, we have used these two very different methods in combination. Our philosophy for CASP3 was to automate the process of structure recognition. Accordingly, we submitted the alignments that were generated automatically by our threading and HMM models. We avoided the temptation to improve these alignments even where it was obvious (by eye) that there were correctable errors. We caved in to a single exception: our last prediction of a cytochrome suggested that there might have been a circular permutation, so in Model 2 we added a short stretch predicting the position of the C-terminal CXXCH motif. Hidden Markov Models (HMMs) First, we constructed a library of HMMs, one for each member of a non- redundant (at the 40% identity level) subset of PDB. Sequence homologs were gathered using FASTA, and aligned using PRRP (Gotoh, JMB 1996). Sequence weighting in HMM construction Because HMM construction using Dirichlet mixture priors to estimate the posterior amino acid distributions in HMMs (Sjolander et al., CABIOS 1996) is quite sensitive to both correlations in the data as well as the number of observations, careful sequence weighting is critical to the effectiveness of the HMMs in remote homolog identification. (If training data contains even a small number of highly correlated sequences, the models are likely to be very specific, but lack sensitivity.) To obtain the relative weights assigned the sequences, we used Henikoff weighting (Henikoff & Henikoff, JMB 1994). To assign the total weights allotted, we estimated the number of independent observations in the data, and chose that number as the total weight. Note: this method departs from the method used in the UCSC-EBI CASP2 HMM library construction (in which Sjolander participated, Karplus et al., Proteins 1997), where the total weight was set to control the information content of the HMM; the information content per column was set uniformly for all HMMs in the library. Subfamily HMM construction One of the outcomes of CASP2, was the realization that building a single HMM for a large and variable subfamily reduced specificity, and resulted in poor alignments in variable regions (such as the specificity loops of 1try, the structure chosen as the template for T0031). This motivated what eventually became a chapter in Kimmen Sjolander's Ph.D. thesis: a method for constructing HMMs for subfamilies that used the information in the family as a whole, where appropriate, but kept statistics separate when necessary. We used a variant of the subfamily HMM construction method described in (Sjolander, ISMB 1997). This method first estimates a phylogenetic tree, and then cuts the tree into subtrees, using a combination of information-theoretic and Bayesian tools (pubs). Each subtree defines a subfamily, and the statistics in each subfamily are combined in a position-specific manner in constructing an HMM for each individual subfamily. To be precise, when we construct an HMM for subfamily s, we add pseudocounts to subfamily s at position p from each subfamily s' (s' != s), proportional to the probability of the amino acids seen at position p in subfamily s', given the posterior Dirichlet mixture density estimated for subfamily s at position p. What this amounts to in practice is that we will use statistics from another subfamily at a position when the other subfamily appears to share common physico-chemical constraints at that position. If the two subfamilies appear to have conflicting physico-chemical constraints, the statistics will be kept separate. In this way, we maintain specificity but can still increase sensitivity. Due to time constraints, in the HMM library construction we constructed general (i.e., whole family) HMMs only. However, subfamily HMMs were constructed for each target family (when possible; some targets had too few homologs, or only very closely related ones, for this task), and for individual folds as needed in discriminating between alternative target-structure matches. Interpreting HMM scores HMM scores are not all equally informative. For an HMM built for one family, a NLL-NULL score of -12 could mean a definite homolog. For another family, a non-homolog, having an entirely different fold, could get the same score. Because of this, we normalized the scores each HMM gave a subset of PDB, obtained Z-scores, and found a cutoff (both in terms of z-score and in terms of NLL-NULL score) for each family where non-homologs came in. Later, when we scored each target against each HMM, we found that in some cases, sorting by z-score put the correct fold to the top of the list (as, for instance, was the case for target T0084, which we identified as a leucine zipper), whereas in other cases sorting by NLL-NULL score produced what we felt was the correct fold. In all cases, we examined potential matches between the target and the top- scoring folds sorted by each method. Protein threading This threading algorithm is a variant of the "3D-1D" profile method first proposed by Bowie & Eisenberg (Science 1991). The protein structure is represented as a linear string of "structural environments." There are nine different types of environment: 3 secondary structure types (helix, strand, other) times 3 degrees of solvent accessibility. This representation has fewer different environment types than used by Bowie & Eisenberg. The target sequence is also represented as two different strings: an amino acid sequence and a sequence of predicted secondary structure for each position. The truly unique part of the algorithm is in the scoring of sequence-structure matches. Rather than using log-odds scores for matches, and arbitrary gap penalties, the scoring problem is cast as an optimization problem (Thomas & Dill, PNAS 1996). Scores for matching target sequence strings with structure strings are determined by an iterative procedure. The target function of the iterative procedure is based on known sequence-structure alignments inferred from structural alignment of distant homologs (sequence identity < 25%). Gap penalties depend on structural environment, and are optimized simultaneously with matching scores. This results in a self-consistent scoring function. An important part of the method is in calculation of structural templates (3D- 1D profiles). A problem with this string representation is in solvent accessibility representation. Accessibilities are calculated in the presence of the entire chain, so deletions of structure positions often make these values inaccurate. We solve this problem by dividing protein structures into domains, and assuming that deletions will affect the representation. The alignment is global-local, i.e. terminal gaps are penalized. For multidomain proteins, we calculate a structural template for each domain, and for the entire protein. For each domain, we calculate two different templates, one for the domain in isolation, and one in the presence of other domains and chains in the PDB structure. This enables us to treat, even if in a simple way, quaternary stabilization effects. This proved to be critical, for example, in identifying the engineered leucine zipper. Combining threading and HMMs For all targets, we used three approaches: (1) standard sequence similarity searching (FASTA, PSI-BLAST) (2) HMMs (3) Threading For all targets, we ran all three methods. All sequences identified as clearly similar to the target by approach (1) were used in approaches (2) and (3). In general, for targets with sequence similarity to a known structure (i.e. a "hit" by approach (1)), the same structure was always identified by the HMMs, and almost always by threading. When the threading and HMMs differed for a strong HMM hit, we submitted the HMM prediction. For cases with little sequence similarity to a known structure, HMMs were, in some cases, able to find a significant match. Again, if these predictions differed from the threading hits, we submitted the HMM prediction. The really interesting cases are, of course, the most difficult ones. For these predictions, HMMs found no significant matches. In these cases, when there was a strong threading hit, we confirmed using HMMs that it was a plausible prediction (among the top weak HMM hits). When neither threading nor HMMs revealed a strong hit, we had two options: to make a speculative prediction or not to predict (i.e. predict a novel fold). We made a speculative prediction if there was a single fold in common among the top threading and HMM predictions. Flow diagram Our basic approach to each target was as follows. 1. Search NR for sequence homologs for each target, using PsiBLAST or iterated FASTA (clustering new sequences found in each iteration into groups, and choosing a representative from each group as a query in the next iteration). 2. Construct an MSA (multiple sequence alignment) from the target and homologs using ClustalW. 3. Construct an HMM from the MSA using modelfromalign. 4. Reestimate the HMM using sequence weighting (as described above). 5. Construct subfamily HMMs. 6. Search PDB with each HMM using local-local alignment (SAM program hmmscore, using swscore 2). 7. Score all the sequence homologs identified in (1) above against the HMM and threading model libraries. 8. Post-hoc analysis: Examine the top-scoring matches found in (6) and (7). If no one match has a high significance (strong z-score or NLL-null score), identify SCOP superfamilies or fold families having overall high scores. Estimate and examine multiple alignments of the target and homologs against putative homologous structures and their homologs. 9. For two of the targets, we submitted 3D coordinates rather than alignments. These were the two examples of high sequence similarity without insertions or deletions. The models were built automatically using two different methods available in the Look and GeneMine products from Molecular Applications Group. The first algorithm is SEGMOD (Levitt, JMB 1992), which allows movement of both backbone and sidechain atoms. The second algorithm, CARA (C. Lee, JMB 1996), fixes the backbone and optimizes only sidechain coordinates. These models were built by Dr. Michael Mueller at MAG.

Trials, Difficulties and Success of Predicting CASP3 Targets

Motonori Ota, Takeshi Kawabata, Akira Kinjo, Ken Nishikawa

We organized the team UNAGI which consists of independent predictors having their own threader. We shared the information about the target, that is, we exchanged information about the sequences or published papers of the experiments, discussed the results of the calculations by our own threaders, motif search, hypothesis for the evolution of the target and the quality of alignments etc., and finally determined our submissions of which we reached agreement. First the target sequences were analyzed with the available tools on WWW: 1) sequence homology search: FASTA, BLAST, PSI-BLAST, and its homologous sequences were aligned by CLUSTALW for the input of threaders; 2) bibliography search from the SWISS-PROT references or PubMed. If there were any structural information such as disulfide-bonds, we tried to confirm its significance by reading the original paper thoroughly. Second the more complicated analysis by open WWW or in-house tools were carried out: 1) secondary structure prediction by SSThread (Ito, Nishikawa), PHD, JOINT (Noguchi, Nishiakwa) and BW-MGOR (Kawabata); 2) motif search by PROSITE and by a program developed by Kawabata which allows more complicated conditions than the former. Third the target sequence, its homologs and their multiple alignment were thread on the PDB representative set. For a single sequence, we performed the threading by COMPASS (Matsuo, Nishikawa), S3 (Kinjo) and LIBRA (Ota). COMPASS uses a set of knowledge-based functions which take into account four physico-chemical terms: side-chain packing, hydration, local-confirmation, and hydrogen-bonding. S3 (Kinjo) uses functions similar to those of COMPASS but more emphasis on local terms (secondary structures and hydrations) than pair-wise long-range terms. LIBRA uses the same terms as COMPASS, but employs different normalization scheme, so it uses the different score table. Also it can accept multiply aligned sequences. If necessary, the inverse folding search, the structure-recognizes-sequence protocol, was carried out with LIBRA. Results of the inverse-folding searches were sometimes useful to confirm those of the forward-folding searches. Sometimes, the inverse-folding searches yielded convincing results even when the forward-folding searches could not give significant results. For the targets whose secondary strucutures were already known, a simple structure alignment program developed by T.Kawabata was performed. The final decision-making was not an easy nor straight-forward process. We could agree immediately on easy targets: our threader yielded the same results and we recognized apparent homology, or multiple results agreed and sequence motif was conserved. Yet most of the time, we could not decide the most appropriate models. In these cases, We customized the programs that took into account the target-specific features such as sequence motifs and disulfide-bonds. In the cases when we could not find any consistency among our results at all, we concluded that the fold was a new one (submitted as NONE). We submitted on 56 models for 25 targets. Among them, 8 structures are determined and publicized by now. We evaluated our submissions and, although further and more thorough assessment is necessary, so far we could say to have succeeded in 4 predictions, namely VanX, polygalacturonase, EH2 domain of EPS15 and MarA protein (1 wrong prediction for Protein HDEA, 3 predictions about which we are not yet sure if we have succeeded or not).

Two Stage Protein Fold Recognition using GenTHREADER and THREADER

Caroline Hadley, Michael Tress and David Jones

For our fold recognition predictions in CASP3, two methods were routinely applied to each target sequence: THREADER (version 2.5) and GenTHREADER (version 3.0). Over the past six years, we have been developing approaches to fold recognition, using a set of statistically determined pairwise potentials, which have proved to be highly effective, and offer considerable scope for future improvement. THREADER 2.5 is the latest incarnation of our threading program, and although it now incorporates a number of new features (including options for locating domains in target sequences), and a more refined set of potentials, the overall concept of the method remains more or less unchanged since CASP2. Firstly, a library of unique, continuous protein domain folds is derived from the database of protein structures. Each fold is considered as a chain tracing through space; the original sequence being ignored completely. The test sequence is then optimally fitted to each library fold (allowing for relative insertions and deletions in loop regions), using a double dynamic programming algorithm, with the 'energy' of each possible fit (or threading) being calculated by summing the proposed pairwise interactions and solvation parameters. The library of folds is then ranked in ascending order of total energy, with the lowest energy fold being taken as the most probable match. Recent features added to the method allow sequence information and predicted secondary structure information to be considered in the fold recognition process. The latest feature is the option to combine threading results for a family of related proteins in order to enhance sensitivity. GenTHREADER is our latest fold recognition method which has been designed to be both fast and reliable, and is particularly aimed at automated genome annotation. The method uses a traditional sequence alignment algorithm to generate alignments which are evaluated by threading techniques. As a final step, each threaded models is evaluated by a neural network in order to produce a single measure of confidence in the proposed prediction. The speed of the method, along with its sensitivity and very low false-positive rate makes it ideal for automatically predicting the structure of all the proteins in a translated bacterial genome. The method has been applied to the genome of Mycoplasma genitalium, and analysis of the results shows that as many as 47% of the proteins derived from the predicted protein coding regions have a significant relationship to a protein of known structure. In some cases, however, only one domain of the protein can be predicted, giving a total coverage of 30% when calculated as a fraction of the number of amino acid residues in the whole proteome. In making CASP3 predictions, GenTHREADER was used as a pre-filter. Where GenTHREADER was able to make a confident prediction (generally in cases where a clear evolutionary link is apparent between the target program and an entry in the fold library), this fold was assumed correct and THREADER was used to generate the final alignment (though with appropriate sequence weighting options). In cases where GenTHREADER did not produce an unambiguous result, full runs were performed with THREADER itself in order to deduce a set of likely folds.

The Generate and Select Ab-inito Prediction Hierarchy: Protein Fold Determination from Sparse Distance Restraints

Derek A. Debe & W. A. Goddard III

We have developed the generate-and-select hirarchy for ab-initio tertiary protein structure prediction. The foundation of this hierarchy is the Restrained Generic Protein (RGP) Direct Monte Carlo method (submitted, J. Chem. Phys.). The RGP method is a highly efficient, off-lattice residue buildup procedure that can quickly generate the complete set of polypeptide topologies that satisfy a very small number of inter-residue distance restraints. For a 100-residue protein with just 4 inter-residue restraints, the RGP method can generate the complete set of topologies (~10^5 structures) in less than one hour using a Silicon Graphics R10000 single processor workstation. Following structure generation by the RGP method, a simple criterion that measures the burial of hydrophobic and hydrophilic residues can reliably select a reduced set of ~10^2 structures that contains the native topology. Next, each of the structures in this reduced set is minimized with repsect to the residue burial function and rescored. This minimization can often rank a native topology structure in the five lowest energy folds. We have developed methods to incorporate secondary structure predictions and sidechains into the folds. The entire prediction hierarchy for a 100-residue protein with 4-10 inter-residue distance restraints can be completed in less than six hours on a single processor workstation. The objective of the generate-and-select hierarchy is to quickly rule out large regions of conformation space based on known inter-residue restraints and hydrophobic criteria. Once large regions of fold space have been eliminated, far more computationally intense methods, such as molecular dynamics or dynamic Monte Carlo can be applied in order to refine the native topology fold. Hence the generate-and-select hierarchy relieves the burden of analyzing large, diverse regions of conformation space with computationally expensive simulation methods. For the CASP3 prediction conference, we used the generate-and-select hierarchy to make predictions on targets T0056, T0059, T0061, T0072, T0075, and T0079. For each target, the initial inputs to the RGP algorithm were a set of 3-8 distance restraints obtained from predictions of disulfide bond connectivity, hydrophobic core residue predictions, or correlated mutation analyses. Sets of on the order of ~10^4 were generated for each target, and the five lowest energy folds were refined to incorporate predicted secondary structure (PHD) and sidechains. For each target, these five lowest energy folds were submitted. Total single-processor computation time for each target was 5-10 hours. Successful application of the generate-and select hierarchy depends on three procedures that are currently of intense interest in the ab-intio structure prediction field: 1. predicting inter-residue contacts by analyzing correlated mutations in distant homologous sequences; 2. efficiently generating folds that satisfy a set of inter-residue distance restraints; 3. recognizing near-native structure from large sets of misfolded candidates. Because of the hierarchical nature of our prediction method, it is possible to understand exactly which stage in the hierarchy is responsible for an inaccurate prediction in instances where the native topology fold is not isolated. Thus, the generate-and-select hierarchy is an excellent vehicle for studying the performance that is required by each of the three procedures to enable accurate low-resolution tertiary structure predictions.

Protein structure prediction with combinatorial optimization

Eckart Bindewald, Ulrike H?er, Matthias Heiler, J?gen Hesser, Reinhard M?ner

A branch and bound algorithm was implemented to search for low energy structures of proteins. The space of possible protein conformations was discretized using a library of building blocks. Each building block represents several amino acids of a protein backbone. A library of 124 building blocks with the length of 6 amino acids was used. This library of building blocks was developed in our group by applying clustering algorithms to a set of fragments of backbone conformations which were taken from protein structures from the PDB database. Protein conformations are represented by the non-overlapping concatenation of building blocks. The branch and bound algorithm searches the energetic minimum in the space of possible building block combinations. The used scoring function is a contact potential. The total energy in this model is the sum of the interaction energies between all its atoms (or pseudo atoms). Interactions between atoms of the same amino acid are not considered. Interactions between atoms of adjacent amino acids are only considered, if they both belong to the side chain. The energy model is based on the atom types and Van der Waals radii described in: Ai-Jun Li, Ruth Nussinov: PROTEINS, 32: 111-127 (1998) . The interaction parameters of the energy model were fitted such that the native structures of 10 test proteins have a lower energy than "good" alternative structures. Predictions were submitted for the Casp-3 targets 59, 61, 65 and 84. They were chosen because among the other targets because they consist of less than 100 amino acids. Possible side chain conformations were taken from a rotamer library given in Ponder, Richards: J.Mol.Biol. 193, 775-791 (1987). Before the start of the conformational search, side chains were placed on the building blocks: For each building block and each possible position on the protein, possible side chain conformations were optimized with a branch and bound algorithm. The library of building blocks was processed, such that for each building block position of the given protein a specialized library of building blocks with side chain information existed. The conformational search was restricted using consensus results from secondary structure predictions given by the JPRED server. The building blocks were classified into 7 different classes, depending on whether they represent a random coil piece or the N-terminal border, C-terminal border or middle part of an alpha helix or a beta sheet. For each building block position the consensus prediction of the JPRED server determined the class of building blocks, which were allowed at this position. The sulfur atoms of the cysteine residues were constrained to have a distance lower than 5 Angstrom in the case of the proteins, which had 2 cysteine residues. The submitted structures correspond to intermediate results of the search algorithm.

Comparative Modelling of CASP3 targets

Nicolas Geux and Mansoor Saqi

Comparative modelling for CASP3 targets was carried out using the Swiss-PdbViewer modelling and visualization tool (SPDBV 3.1). The same general procedure was followed for each model, namely each target sequence was aligned onto its best template (the one with the maximum identity with the target); the alignment was adjusted manually, with particular consideration to the placing of gaps; whenever possible, additional templates were superimposed onto the best template in hope of adding information about sidechain position; then, an average model was built with SWISS-MODEL. Geometrically possible loops were generated de novo directly within Swiss-PdbViewer by a combinatorial approach and inspected manually. Those with favourable H-bonding network and acceptable polarity of exposed sidechains were retained. Finally, all models were energy minimized. In the case of t0069, the best template was a trimer. Sidechains present at the interface of the three monomers were optimised by performing a combinatorial search of sidechain rotamers. When too many sidechains were interacting with each other to perform an exhaustive search, a simulated annealing procedure was applied.

Threading score functions without Boltzmann statistics

Thomas Huber, Daniel Ayers, Anthony Russell, Andrew E. Torda ANU supercomputing Facility and Research School of Chemistry, Australian National University Canberra ACT 0200 Australia

Our methodology used sequence to structure threading, but with unusual score functions and three distinct computational steps: 1. alignment of the sequence to each structure in a library of templates using a purpose built score function / force field 2. ranking of model structures using a second score function 3. sidechain placement using a third force field and iterative mean field approach For the first step (sequence to structure alignment) we used a score function which allowed direct use of a Needleman and Wunsch algorithm, without relying on the frozen approximation or double dynamic programming. This score function is quite unique in using the identity of only one member of each interaction pair. This means that one can truly score each residue at each possible template position without further approximation. It allows one to make the outrageous claim of a guaranteed optimal sequence to structure aligment, with the caveat of a distinctly non-optimal force field. The second step was the ranking of alignments (models). A second score function was used with the original approximations removed and using the identity of both members of each interaction pair. Finally, in a third step, we showed unbounded faith in our ranked models by trying to place side-chain atoms on the backbones. This was done with a mean-field approach and a more conventional atomistic force field. Side-chains were placed according to a rotamer library and assigned probabilities according to their energies. Rotamer populations were then recalculated in the field due to their neighbours and iterated to self consistency. For the first and second steps, score functions were built without the reliance on Boltzmann statistics used in many other "knowledge-based" force fields. Instead, our goals were described and translated into a cost function whose parameters were to be optimised. In this case, the cost function was based on the ability to distinguish a library of native folds from a very large (10 000 000) number of misfolded structures. Thus, we required that the score functions discriminate good from bad sequence-structure pairs. There was no requirement that the functions have any relation to a physical energy. The method was applied for the second force field, but could also be used for the first force field which only used the identity of one member of each interaction pair. The score functions were low-resolution force fields with 5 or less interaction sites per residue and all of the residue's identity carried by a site on the beta carbon. Separate interaction parameters were used for short, medium and long range sequence separation and were usually based on a hyperbolic tan (sigmoidal) functional form. Although they were not used for CASP3, we have tried several variations of score functions. The optimisation methods have been used to build table-driven force fields which have very similar performance and we have experimented with slightly fewer interaction sites per residue. Since CASP3, we have begun to work on methods to produce more specialised force fields, optimised specifically for properties such as alignment capability.

Protein Structure-sequence threading using mean force potential with 6 degrees of freedom

Kentaro ONIZUKA, Tamotsu Noguchi, Yutaka Akiyama

Multiple dimensional distribution is represented with fewer number of parameters by linearly expanding the distribution and controlling the cut-off orders of expansion. We adopted this method to the distribution of the relative position between two amino-residues in a protein chain, and applied it to the protein fold recognition problem. We compared the recognition ratio of three cases, adopting the distribution 1) with respect to the distance (one degree of freedom) , 2) with respect to the 3D position (three degrees of freedom), and 3) with respect to the 3D position and the relative orientation (six degrees of freedom). The result is that the self-recognition ratio of multiple dimensional distribution is far better than that of the conventional distribution with respect only to the relative distance.

Protein Structure Prediction Using Potentials Derived from the Knowledge-Based Database RELOR

Alfons Haedener

A Knowledge-Based Database of Parameterised Interactions of Amino Acids in Proteins Of three-dimensional structures of proteins, available via the Protein Data Bank (PDB) [1] a representative subset [2] was used to construct a database containing records associated with pairs of amino acid residues of a particular protein of the subset. For the parameterisation, each residue pair of a protein was considered separately and gave rise to one record of the database. A record basically comprises six floating-point numbers which, as an ensemble, relate to the interactions of the pair of residues being considered. The database, called RELOR, currently contains 1'529'378 records derived from 53'915 residues of 267 polypeptide chains. Another part of the RELOR database contains 658'686 records, each associated with the interactions of an individual amino acid residue with an ion or a particular molecule that is not part of the polypeptide chain, such as a metal ion, a water molecule of the ordered solvent structure, or a cofactor (A. Haedener, unpublished). Potentials for Protein Conformations The database RELOR was used to derive a potentials that are useful in calculations of the relative stability of given conformations of a protein. Basically, a potential is evaluated by considering individual residue pairs of a given conformation of the test protein, followed by estimating the frequency of the occurrence of similar residue pairs in the database. This evaluation generates a number that can be related to the relative stability of the given conformation. The method can be validated by checking the output number against empirical data describing the relative stabilities of particular proteins and their site-directed mutant variants. Such empirical data has been generated for some well-studied test proteins such as barnase [3] or bacteriophage-T4 lysozyme [4]. The software that creates the potentials, called PROFITUFT, is written in C and runs on a Silicon Graphics O2 Workstation. Both RELOR and PROFITUFT are continuously under development. In principle, interactions of a protein with the ordered water structure or a contingent cofactor, as encoded by RELOR, could be taken into consideration by PROFITUFT. This is not currently implemented in the program, however. Simulation of Folding Processes Potentials generated by PROFITUFT can be integrated into methods of all major fields of protein structure prediction, i. e. homology modelling, threading, and, in particular, ab-initio methods [5]. The algorithm for the submitted ab-initio prediction for target T0084 tries to simulate the folding process of a protein. It is based on methods using simulated annealing in both discrete and continuous conformational space [6] [7] in which the dihedrals phi and psi of individual residues are independent variables. Due to a number of restrictions built into the current version of the algorithm, the predicted model of T0084 is thought to merely represent a kind of molten-globule state of the polypeptide chain. References [1] F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, M. Tasumi, J. Mol. Biol. 1977, 112, 535ㄽ542. [2] U. Hobohm, M. Scharf, R. Schneider, C. Sander, Prot. Sci. 1992, 1, 409ㄽ417. [3] A. R. Fersht, FEBS Lett. 1993, 325, 5ㄽ16. [4] B. W. Matthews, Annu. Rev. Biochem. 1993, 62, 139ㄽ160. [5] J. Moult, J. T. Pedersen, R. Judson, K. Fidelis, Proteins: Struct. Funct. Genet. 1995, 23, issue no. 3. [6] W. H. Press, B. P. Flannery, S. A. Teukolsky, W. T. Vetterling, 'Numerical Recipes in C', 2nd edition, Cambridge University Press, Cambridge, 1992. [7] J. A. Nelder, R. Mead, Computer Journal 1965, 7, 308.

fold recognition for casp3

daniel fischer

I have submitted predictions to most of the available targets for casp3, including those with homology to known structures. The latter submissions are interesting to assess the alignment accuracy. The predictions were based on the SDP method of Fischer & Eisenberg as implemented in the fold recognition server frsvr. As far as possible, I attempted to submit the automatic results of the server with no manual intervention. No biological/literature information was used to bias the predictions. When a target had homologous sequences in the database, independent runs for each of the homologous sequences was attempted. When similar results were obtained for the homologous sequences, the confidence of the prediction increased. When the prediction scores were low and/or inconsistent, I suspected of a possible novel fold. The main problem encountered was when the prediction seemed to be weak and one fold needed to be selected. In such cases a rather arbitrary decision was taken

Methodology and strategy for comparative modeling

Lisa Yan, David J Edwards

This abstract outlines the methods and our strategy for predicting protein structures using comparative modeling method. Among the list of CASP3 target sequences, 14 of them are marked as having homologous sequences of known structures. We entered predictions for 10 such sequences. Searching PDB database using FASTA or BLAST program identifies the template structures used to create homology models. Models are built using either Modeler4 or Modeler5 program based on all possible templates together or separately. Since the accuracy of sequence alignment is a crucial part in homology modeling, we tried to create sequence alignment using various alignment methods and parameters. If there are multiple template structures available, the first step is to align the template sequences based on their structure similarity or sequence similarity. Multiple sequence alignment based on structure similarity is generated by comparing the C alpha distance matrix using the automatic divide and conquer algorithm in InsightII. Multiple sequence alignment based on sequence similarity is created using ClustalW or divide and conquer algorithm in InsightII. Then, the model sequence is aligned to template(s) using pairwise global alignment algorithm in InsightII or using structure enhanced global alignment algorithm (Align2D) provided by Modeler. We also tried to align all sequences, including templates and model, in one step using multiple sequence alignment method from ClustalW. Based on the alignments generated by various methods, models are created using either Modeler4 or Modeler5. For some models, loop regions are refined using statistical pair potential in Modeler5. The loop regions are defined automatically by Modeler5 as the sequence segments not aligned to any templates. Models are then selected according to model evaluation score calculated by Profile-3D/verify in InsightII. We found that the models created based on the alignments generated using Align2D method usually give better verify scores.

Threading score functions without Boltzmann statistics

Thomas Huber, Daniel Ayers, Anthony Russell, Andrew E. Torda ANU supercomputing Facility and Research School of Chemistry, Australian National University Canberra ACT 0200 Australia

Our methodology used sequence to structure threading, but with unusual score functions and three distinct computational steps: 1. alignment of the sequence to each structure in a library of templates using a purpose built score function / force field 2. ranking of model structures using a second score function 3. sidechain placement using a third force field and iterative mean field approach For the first step (sequence to structure alignment) we used a score function which allowed direct use of a Needleman and Wunsch algorithm, without relying on the frozen approximation or double dynamic programming. This score function is quite unique in using the identity of only one member of each interaction pair. This means that one can truly score each residue at each possible template position without further approximation. It allows one to make the outrageous claim of a guaranteed optimal sequence to structure aligment, with the caveat of a distinctly non-optimal force field. The second step was the ranking of alignments (models). A second score function was used with the original approximations removed and using the identity of both members of each interaction pair. Finally, in a third step, we showed unbounded faith in our ranked models by trying to place side-chain atoms on the backbones. This was done with a mean-field approach and a more conventional atomistic force field. Side-chains were placed according to a rotamer library and assigned probabilities according to their energies. Rotamer populations were then recalculated in the field due to their neighbours and iterated to self consistency. For the first and second steps, score functions were built without the reliance on Boltzmann statistics used in many other "knowledge-based" force fields. Instead, our goals were described and translated into a cost function whose parameters were to be optimised. In this case, the cost function was based on the ability to distinguish a library of native folds from a very large (10 000 000) number of misfolded structures. Thus, we required that the score functions discriminate good from bad sequence-structure pairs. There was no requirement that the functions have any relation to a physical energy. The method was applied for the second force field, but could also be used for the first force field which only used the identity of one member of each interaction pair. The score functions were low-resolution force fields with 5 or less interaction sites per residue and all of the residue's identity carried by a site on the beta carbon. Separate interaction parameters were used for short, medium and long range sequence separation and were usually based on a hyperbolic tan (sigmoidal) functional form. Although they were not used for CASP3, we have tried several variations of score functions. The optimisation methods have been used to build table-driven force fields which have very similar performance and we have experimented with slightly fewer interaction sites per residue. Since CASP3, we have begun to work on methods to produce more specialised force fields, optimised specifically for properties such as alignment capability.

Post-processing of Secondary structure prediction by using reverse protein sequences.

Jong H. Park and George M. Church

Present protein secondary structure prediction programs rely on training or utilizing secondary structure assignment information from 3D structures. The highest average accuracy of them is below 75%. There are various post-processing approaches to extend the accuracy of such prediction methods. Most of them combine different prediction algorithms and produce consensus predictions with various intuitive rules. If any target sequence has many homologous sequences known, the secondary structure prediction programs can utilize multiple sequence alignments resulting in relatively higher prediction accuracy. It can also be useful to analyze how difficult a prediction can be, especially when there are few homologous sequences for the target. CASP3 can provide people with independent test protein sequences for such post processing programs as the target sequences chosen are pre-filtered and categorized into sequences with known and unknown structures resulting in only relatively unique protein targets. We have tested a post processing approach that utilizes the reverse protein sequences to see 1) how unreliable the predicted sequence segments can be 2) how we can improve the accuracy of the predictions made by major prediction algorithms. The method is based on previous empirical observation that strongly predicted secondary structures of for- and backward sequences of the same protein tend to be similar and often identical. The program developed in perl programming language, ㄣpredict_ssop', accepts the output of PHD (Rost and Sander, 1993) and Predator (Frishman and Argos, 1997) programs. It first runs either or both of the 2 programs with a sequence or a homologous sequence set. Then it reverses the given sequence(s) and runs the programs again. It, then, reverses back the predictions for the reversed sequences. With both forward and backward secondary structure predictions it compares the prediction made for each position while rewarding and penalizing the prediction according to the identity of the comparison. For example, with a given single sequence, if both forward and backward sequences showed alpha helix prediction for a certain residue position, the score is the average multiplied by incompletely optimized positive factor of the reliability of the predictions from either PHD or Predator. If forward and backward predictions do not match for a residue position, penalty is given to the prediction reliability lowering the original reliability. When helix and beta sheet prediction with high reliability is matched to a no prediction or coil prediction with low reliability, helix and beta prediction is not penalized. If a multiple sequence alignment is given there can be two ways to deal with the prediction. One is to run a prediction program for each sequence in the alignment and give reliability scores in comparison with all the corresponding backward sequences of the alignment with the same rule described above. The other way is to run the prediction program over the multiple sequence alignment and calculate the reliability score from the back- and forward sequences of the target sequence in the alignment. At the assessment of the prediction competition, we hope to be able to analyze the result to know 1) if there is any biological reason for a reversed sequence has the same or similar secondary structure prediction 2) how to utilize the control information given by the reverse sequence in post-processing prediction outputs 3) how to improve the prediction algorithms to integrate the control reverse sequence in the stage of neural net training (for PHD like program) and propensity table generation (most other programs). Acknowledgement J.P. is supported by a grant from Hoechst Marion Roussel. References Rost B, Sander C., Proc Natl Acad Sci U S A 1993 Aug 15;90(16):7558-62 Improved prediction of protein secondary structure by use of sequence profiles and neural networks. Frishman D, Argos P., Proteins 1997 Mar;27(3):329-35 Seventy-five percent accuracy in protein secondary structure prediction.

Evolution-based Transparent Structure Prediction

Steven A. Benner, Gina Cannarozzi, Dietlind L. Gerloff

The use of alignments of multiple protein sequences to predict the conformation of proteins, first proposed in 1977 by Lenstra et al. (Lenstra, J. A., Hofsteenge, J. & Beintema, J. J. (1977). Invariant features of the stuructre of pancreatic ribonuclease. J. Mol. Biol. 109, 185-193) has become commonplace since the approach generated a "remarkably accurate" prediction for protein kinase in 1991 (Benner, S. A. & Gerloff, D. (1991). Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: The catalytic domain of protein kinases. Adv. Enzyme Regulat. 31, 121-181). This prediction provided the first example where structure prediction was used to deny long distance homology implied by limited sequence similarities, an example of "compensatory covariation analysis" in structure prediction, and an early example of functional genomics based on evolutionary analysis. Despite numerous papers (reviewed in Benner, S. A., Cannarozzi, G., Chelvanayagam, G., Turcotte, M. (1997) Bona fide predictions of protein secondary structure using transparent analyses of multiple sequence alignments. Chem. Rev. 97, 2725-2843) showing the importance of higher order "non-Markovian" analysis of the divergence of protein sequences, rigorous evolutionary modelling, and human involvement in the process by which analytical tools are developed, the overwhelming majority of structure predictors continue to rely on Markov models, first order analyses, neural nets, and fully automated computational tools to predict protein structure and analyze its implications for function. This poster will make another attempt to persuade our computational friends and colleagues to think "outside the box" when using sequence data to analyze and predict protein structure, and to apply it to functional genomics.

Monte Carlo Threading

Leonid Mirny and Eugene Shakhnovich

The energy of a protein conformation is computed using a contact pairwise approximation. Two residues are said to be in contact if the distance between their Cb atoms is less than 8A. Conformation of the lowest energy is found by threading of the target sequence through known protein structures. The major features of our method are the following. * Parameters of the energy function (potential) are obtained by an optimization procedure, which minimizes the Z scores of the proteins in the training set. The Z score is the deviation of the energy of the native conformation from the average energy of compact random conformations, measured in the units of standard deviation. To minimize the Z scores simultaneously for all proteins in the training set we minimize the harmonic mean of individual Z scores. For a given form of energy function, this procedure yields a potential which provides minimal Z scores. We extensively studied convergence and generalization of the method. * In threading, the target sequence is used together with its close homologs. Sequences homologous to the target are found by a standard sequence alignment procedure. For each homologous sequence (m) we compute a matrix of interactions B^(m). An element Bij^(m) of this matrix gives the energy of a contact between residues in positions i and j of sequence (m). Then we compute an average matrix of interactions Bij= where averaging is done over all homologous sequences and the target sequence. This matrix is used in the threading procedure to compute the energy of a protein conformation. * We start from fast fold recognition procedure which makes threading through every structure in a non-redundant database (about 1000 structures). Fast fold recognition does no relay on the frozen approximation, instead, for each structure it enumerates all possible sequence-structure alignments containing a single gap. Finding a sequence-structure alignment of the lowest energy in a general case (any number of gaps) is computationally expensive, therefore we first select best scoring structures using a fast but approximate fold recognition and then subject these structures to the exact algorithm. Fifty structures which provide the lowest Z scores are selected for detailed threading. * Monte Carlo (MC) threading is the core of our prediction algorithm. MC threading is able to find a sequence-structure alignment of the minimal energy in a general case without relaying on frozen approximation or any a-priory knowledge of aligned fragments. No gap penalties are used and the whole process of threading is driven by energy function alone. To avoid too short fragments in the alignment we constrain the minimal length of a fragment to 6 residues. MC threading starts from a random sequence-structure alignment. At each step it makes a small change in the alignment (moves/shrinks/extends or splits a fragment) and then accepts of rejects this move according to the energy change and Metropolis criteria. Simulated annealing protocol is used to find an alignment of the lowest energy. Importantly, this algorithm also allows sampling of sub-optimal alignments. A structure and an alignment with the lowest Z score constitutes a prediction. Alternately, contact maps obtained from 20 structures with the lowest Z score are clustered into groups of similar maps and the average map of the largest cluster is used as the predicted contact map. REF: Mirny, L.A and Shakhnovich, E.I. "How to determine protein folding potential? A new approach to the old problem." J.Mol.Biol. 1996, v264, p1164-1169 Mirny, L.A. and Shakhnovich, E.I. "Fold Recognition by Threading: Why it Works, Why it Doesn't" J.Mol.Biol. 1998, v283, p507-526

Iterative sequence search method using intermediate sequences to detect very distant sequence homologs for fold recognition.

Jong H. Park

It has been shown that iterative sequence search method with multiple intermediate sequences can increase the sensitivity of popular sequence search method by 2 folds (http://cyrah.med.harvard.edu/Proj/Bio/Search_meth_comp/assess_final.html, Park, et al., in press). Iterative search method using Hidden Markov Model (HMM) has shown to be marginally more sensitive than position specific matrix blast algorithm (PSI-blast, Altschul, et al). However, PSI-blast is simpler and easier to do the iteration for the moment. Optimal evalue has been sought for PSI-blast using already known structural database (PDB40D) based on structural classification of proteins (SCOP). Utilizing the optimally estimated evalue for PSI-blast as well as HMM and manual sequence search, fold recognition was carried out for CASP3. The assessment result will show the extent of the latest sequence search method which uses multiple sequence information with iteration in comparison with threading algorithms. The search was done against non-redundant sequence search database which is used as a sequence pool for intermediate sequences. Each prediction target was subject to automated perl script which runs PSI-blast algorithm with the evalue of 0.001 which is known to be the optimal value at the error range of 1%. Iteration was done up to 20 times unless conversion occurred. At each stage of iteration, HMM (HMMER beta version 2.0) was built to search the non-redundant and PDB100D database which contains all the domain sequence for known protein structures. Also, manual checking was carried out to identify functionally similar matches in the iteration process. Sequence alignments which belonged to boundary regions of evalue (close to 0.001 in both ways) were checked manually at each iteration. With extremely unique sequence targets, all the hits of the initial PSI-blast were examined for any distant functional and alignmental links. This comparison experiment of latest pure sequence search method against threading algorithm with unknown prediction targets of Casp3 will provide us a valuable insight on how well the latest sequence search methods compare with threading, how different in their ranked hits and knowledge on future improvements on sequence search algorithms. Acknowledgement J.P. is supported by a grant from Hoechst Marion Roussel. References Altschul, S.F., Madden, T.L., Sch,,ffer, 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 Acid Res. 25, 3389-3402. Jong Park, Kevin Karplus, Christian Barrett, Richard Hughey, David Haussler, Tim Hubbard and Cyrus Chothia, Sequence Comparisons Using Multiple Sequences Detect Twice as Many Remote Homologues as Pairwise Methods, JMB, in press.

The protein secondary structure prediction by New Joint Method and the SSThread method.

Tamotsu Noguchi

The joint prediction method for protein secondary structure combines the result of different prediction methods. In the original New Joint Method, the five best methods (QS: Qian-Sejnowski, Na: Nagano, PF: Ptitsyn-Finkelstein, NO: Nisikawa-Ooi, GGR: Gibrat-Garinier-Robson) of eight examined, are chosen, each of which represents a different methodology (Nishikawa and Noguchi, 1991). The average prediction accuracy of a test set was 64.8 % in the three-state assessment of alpha helix, beta sheet and coil. In the New Joint Method at present, the Na method has been replaced by a method for the prediction of protein secondary structure using the 3D-1D compatibility algorithm (SSThread method), which have been proposed by Ito et al.. The New Joint Method is available through the protein secondary structure prediction menu on the PAPIA (PArallel Protein Information Analysis) WWW. The PAPIA WWW page can accessed by the URL (http://www.rwcp.or.jp/papia/papia.html). The structural library of the 3D-1D compatibility search used in the SSThread method contains 325 protein chains, which have been selected from PDB release 69 by Matsuo and Nishikawa (1994). The average prediction accuracy by the SSThread method is 70.3 %. Though the New Joint Method replaced from the Na method to the SSThread method have improved the prediction accuracy from 64.8% to 66.6 %, the SSThread method is better than the New Joint Method. However, the prediction accuracy for the SSThread method drops in the case of no similar protein in the structural library. Therefore, we prepare the SSThread using the new structural library, which contains 817 protein chains selected from PDB release 83 besides the SSThread used in the New Joint Method. We have predicted the protein secondary structures of CASP3 target proteins by the New Joint method and the SSthread method using the new structural library. The result of the prediction of protein secondary structure have been based on the results of the New Joint Method, in case the target size is smaller than 100 residues or a similar protein with a target, whose total compatibility score is better than -2.8, is not found in the new structural library by the 3D-1D compatibility search. We have predicted the secondary structure of the other target proteins by the SSthread method using the new structural library, since the SSthread method is better than the New Joint Method in that case. In this case, we have aligned the target protein with the highest score protein according to the 3D-1D alignment and submitted in format to express unambiguous alignment to PDB entries.

Comparative protein structure modeling by MODELLER-5

Andras Fiser, Roberto Sanchez, Francisco Melo, Azat Badretdinov and Andrej Sali

Template selection: Templates were identified by programs PSI-BLAST (1), ALIGN (2), PROFIT (3), UCLA-DOE fold assignment server (4), and MODELLER (5,6). In general, several different template combinations were used, finally picking those templates that resulted in the best model, as evaluated by the program PROSAII (7). Alignment: Alignments were generated with programs PSI-BLAST (1), ALIGN (2), MODELLER (5,6), ITERALIGN (8), and CLUSTALW (9). Template structures were aligned by the least-squares rigid body superposition in MODELLER-5. When possible, sequences from GENPEPT that were easily alignable with the template structures were added to the alignment, as were those sequences that were easily alignable to the target sequence. Multiple sequence alignments (ITERALIGN) as well as profile to profile sequence alignments were performed (CLUSTALW, MODELLER-5). In many cases, the alignment was also edited by hand taking into account the template structures, predicted secondary structure, functional residues, etc. Among the many generated alignments, the final alignment was that which resulted in the best model, as evaluted by the program PROSAII (7). Model building: Model building was done with MODELLER-5 (5,6). The input to the program was the alignment between the target sequence and the template structures. The output, obtained without any user intervention, was a model with all non-hydrogen atoms. When possible, additional restraints originating from ligand binding and quaternary interactions were used. Loops were modeled subsequently with a new automated procedure in MODELLER-5 (A. Fiser and A. Sali, in preparation). Model evaluation: All models, except one, were checked with PROCHECK (10), PROSAII (7), and WHAT-CHECK (11). 1) S.F. Altschul et al. Nucl.Acid.Res. 25, 3389, 1997. 2) S.F. Altschul. Proteins 32, 88, 1998. 3) H. Flockner et al. Proteins 23, 376, 1995. 4) http://www.mbi.ucla.edu/people/frsvr/preds/MG/MG.html 5) A. Sali and T.L. Blundell. J.Mol.Biol. 234 779, 1993. 6) http://guitar.rockefeller.edu/ 7) M.J. Sippl. Proteins 17, 355, 1993. 8) L. Brocchieri and 11) S. Karlin. J. Mol.Biol. 276, 249, 1998. 9) D.G. Higgins and P.M. Sharp. CABIOS 5, 151, 1989. 10) R.A. Laskowski et al. J.Appl.Cryst. 26, 283, 1993. 11) R.W.W. Hooft et al. Nature 381, 272, 1996.

The protein secondary structure prediction by New Joint Method and the SSThread method.

Tamotsu Noguchi, Kentaro Onizuka, Yutaka Akiyama

The joint prediction method for protein secondary structure combines the result of different prediction methods. In the original New Joint Method, the five best methods (QS: Qian-Sejnowski, Na: Nagano, PF: Ptitsyn-Finkelstein, NO: Nisikawa-Ooi, GGR: Gibrat-Garinier-Robson) of eight examined, are chosen, each of which represents a different methodology (Nishikawa and Noguchi, 1991). The average prediction accuracy of a test set was 64.8 % in the three-state assessment of alpha helix, beta sheet and coil. In the New Joint Method at present, the Na method has been replaced by a method for the prediction of protein secondary structure using the 3D-1D compatibility algorithm (SSThread method), which have been proposed by Ito et al.. The New Joint Method is available through the protein secondary structure prediction menu on the PAPIA (PArallel Protein Information Analysis) WWW. The PAPIA WWW page can accessed by the URL (http://www.rwcp.or.jp/papia/papia.html). The structural library of the 3D-1D compatibility search used in the SSThread method contains 325 protein chains, which have been selected from PDB release 69 by Matsuo and Nishikawa (1994). The average prediction accuracy by the SSThread method is 70.3 %. Though the New Joint Method replaced from the Na method to the SSThread method have improved the prediction accuracy from 64.8% to 66.6 %, the SSThread method is better than the New Joint Method. However, the prediction accuracy for the SSThread method drops in the case of no similar protein in the structural library. Therefore, we prepare the SSThread using the new structural library, which contains 817 protein chains selected from PDB release 83 besides the SSThread used in the New Joint Method. We have predicted the protein secondary structures of CASP3 target proteins by the New Joint method and the SSthread method using the new structural library. The result of the prediction of protein secondary structure have been based on the results of the New Joint Method, in case the target size is smaller than 100 residues or a similar protein with a target, whose total compatibility score is better than -2.8, is not found in the new structural library by the 3D-1D compatibility search. We have predicted the secondary structure of the other target proteins by the SSthread method using the new structural library, since the SSthread method is better than the New Joint Method in that case. In this case, we have aligned the target protein with the highest score protein according to the 3D-1D alignment and submitted in format to express unambiguous alignment to PDB entries.

Structure prediction of CASP targets T0056, T0059, T0065 and T0084.

Chen Keasar and Michael Levitt

We predicted the structures of four target proteins: T0056, T0059, T0065 and T0084. The conformation space of each protein was sampled as well as those of distant but clearly related homologous proteins (if available). The submitted models were low in energy, compatible with the homologous proteins and different from one another. As a major guideline for the conformation space sampling of T0059, T0065 and T0084, we used consensus secondary structure prediction. For T0056, we used the available secondary structure. The conformation space of all given proteins were sampled by selecting a large number (>= 10,000) of random (but rather extended) structures with the predicted secondary structure. These structures were then minimized with the program ENCAD using the va09a algorithm in torsion angle space. The torsion angles of residues with assigned secondary structures were forced to remain close to the ideal value during the minimization. The energy function used is a combination of the united atoms ENCAD forcefield and heuristic terms for hydrogen bonds hydrophobicity and hydrophilicity. These heuristic terms were developed in an iterative process. First, the conformation space of ten small proteins (1bba, fc2, 1gpt, 2ovo, 4pti, 1shf, 1igd, 2cro, 1ctf and 1ubi) were sampled with the above procedure, using the native secondary structure. Then the heuristic terms were modified in order to penalize the most bizarre low energy conformations and the conformation spaces were sampled again. Although sometimes these modifications resulted in unexpected artifacts, the process gradually increased both the fraction of native-like structures in the samples and the correlation between the RMS deviation from the native structure and the energy. A major challenge in this process was the generation of beta sheets. With only a simple two-atom hydrogen bond term, the collapse of the protein to a compact structure stopped after the generation of few random hydrogen bonds that fixed the structure. The current term thus includes a cooperative four-atom terms that favors the characteristic hydrogen bond patterns of beta sheets and alpha helices. Less common patterns are penalized. With this type of hydrogen bond term, we are able to generate protein like parallel and anti parallel beta sheets. The hydrophilicity term is penalizes contacts between hydrophobic (aliphatic and aromatic carbon) atoms to charged ones. This is also a cooperative term. For each charged atom, the energetic price grows exponentially with the number of hydrophobic atoms that contact it. Thus, charges on the surface (with few contacts) are hardly penalized but buried ones (with many contacts) are penalized heavily. In contrast, he hydrophobic term is non-cooperative. It pushes each pair of hydrophobic atoms towards one another. This term grows very slowly (log) with the distance. This way even in the beginning of each simulation, when the structures have a very large radius of gyration, the hydrophobicity does not override the other energy terms. At the time of writing this abstract we do not know yet how accurate our predictions are. Our expectations are rather modest. First, the correlation between rms and energy is still rather low and second, the results depend heavily on the quality of the secondary structure predictions. We do believe however that we present a novel approach to protein structure prediction that can be developed into a useful tool.

Modeling by (possibly remote) homology using PSI-BLAST and SCWRL

Roland L. Dunbrack, Jr.

We base our protein structure prediction on finding proteins homologous to the target sequence within the Protein Databank. That is, we are not looking for ﾓ analogous ﾔ folds -- proteins which share the same fold but are not related by evolution to the target sequence. And we are not performing long simulations, which at this time are not likely to succeed in any significant fraction of cases to be useful in real situations of interest in biology. To find homologues in the PDB of each CASP3 target sequence, we first used PSI-BLAST (downloaded from NCBI and run locally) to find homologues of the target sequence in the non-redundant (ﾓ nr ﾔ) GenBank protein sequence database. PSI-BLAST was run through four iterations, using an E-value cutoff of 0.0001 on each round to construct the position-specific matrix for each subsequent round. The final matrix was saved and used to search the PDB sequence database. We also used ﾓ intermediate sequences ﾔ to establish relationships between the target sequence and a PDB sequence. This was accomplished by using each relative of the target sequence as a query sequence against the PDB. Alternatively, we also constructed a database of all sequences in GenBank related to PDB sequences and used this database as the target of BLAST searches. Both of these methods are in fact more sensitive than using the profile constructed by PSI-BLAST, but more time-consuming. When the target sequence was defined by the experimentalists as homologous to a PDB sequence, this method provided the identity of the homologous sequence and a preliminary sequence alignment. When the target was supposedly not related to a sequence in the PDB, this method was used for ﾓ fold recognition ﾔ -- in some cases finding a sequence very distantly related to the target sequence and providing a reasonable alignment. In both situations, the alignments were adjusted manually by inspection of the PDB structure and the positions of insertions and deletions in the target-PDB alignment provided by BLAST. In most cases, these insertions and deletions were already in loop regions and looked quite reasonable. In a few cases, these occurred in regular secondary structure, usually in regions of the alignment that were not highly conserved. The alignments were adjusted to move these insertion or deletion into nearby loop regions. Once an alignment was obtained, we built the backbone by copying the backbone coordinates for all residues from the PDB sequence aligned to any residue in the target sequence. Sidechains were built with the program SCWRL (Bower M, Cohen F, & Dunbrack RL, J. Mol. Biol. 267, 1268 (1997)), keeping sidechain conformations of non-mutated residues, and using SCWRL to build all mutated sidechains. Sidechain conformations were taken from the backbone-dependent rotamer library of Dunbrack. The most recent version of SCWRL (2.1) achieves a prediction rate of 80% correct chi1 angles in self-backbone tests on 316 proteins. SCWRL uses an energy function consisting of log probabilities from the backbone-dependent rotamer library for the local backbone-sidechain interaction and a simple steric function for non-local sidechain-backbone and sidechain-sidechain interactions. Given the approximate nature of the backbone model for homologous targets, we did not refine or perform energy minimization on the sidechain conformations predicted by SCWRL.

Predictor Assessment of Four de novo Tertiary Structure Predictions, Three Motif-Assisted Tertiary Structure Models and Three Threading Alignments Generated through Transparent Analysis of the Patterns of Evolutionary Divergence in Homologous Sequences

Dietlind L. Gerloff, Gina M. Cannarozzi, Marcin Joachimiak, Fred E. Cohen, and Steven A. Benner

In a collaborative effort to explore the potential for predicting supersecondary/tertiary structure in the transparent secondary structure prediction method developed by some of us [1], predictions were submitted for ten CASP3 target proteins lacking significant overall sequence similarities with previously known protein structures. Most importantly, using transparent approaches for secondary and supersecondary/tertiary structure prediction allows us to rationalize why prediction mistakes occurred where they did, and to refine our heuristics based on these insights.Target selection was based primarily on the availability of homologous protein sequences in adequate numbers and evolutionary distributions for the particular problems and for the approaches we chose to apply in each case. As such, we have divided our submissions in three groups, reflecting the differences in starting information and, accordingly, the particular structural aspects we attempted to predict. 1. De novo targets: T0043 HPPK, T0052 CN-V, T0063 IF5A, T0075 ETS-1. Because of the absence of indications for known folds for these targets, predictions for T0043, T0063, and T0075 were approached "ab initio", according to the guidelines described in [1]. Secondary structures were predicted from manually refined multiple sequence aligments [2] based on periodicity in 'SURFACE', 'INTERIOR' and 'ACTIVE SITE' assignments through evolutionary sequence analysis heuristics. In general, plausible supersecondary and tertiary structures were then predicted semi-systematically from indications regarding the relative orientation of secondary structural elements, other indications (e.g. minimum connection lengths between elements), and functional considerations. Where possible, we performed combinatorial analyses of plausible topologies to predict low resolution tertiary structures. Encouraged by our CASP2-results for Hsp90 (T0011) [3], we put particular emphasis on modeling the active site arrangement for T0043, a pyrophosphokinase, based on a predicted mechanism. The active site model narrowed down the number of preferred topologies to six (only one of which could be modeled for CASP due to time constraints). For non-catalytic targets in this group models had to be based on known structures, with suspected similarities in supersecondary structures, due to insufficient distance constraints. A prediction for T0052 was submitted as a modeling excercise of a target lacking sufficient homologous sequences, mostly to reflect our prediction of all-beta secondary structure based on an internal sequence repeat. 2. Targets with sequence motifs containing structural information: T0054 VANX, T0074 EPS15, T0079 MARA. For these targets, the core folding topologies could be conjectured, partially, based on the occurrence of PROSITE [4] or unlisted sequence motifs/signatures including key functional and structural residues. While this information facilitated tertiary structure prediction, subsignificant scores by publicly accessible fold recognition programs (see below) indicated significant structural divergence between the target and possible fold templates (major insertions/deletions, differences in orientation between the modules of bipartite structures, etc.). In our submissions, we attempted to identify peculiarities in the multiple sequence alignments of the targets vs. templates (e.g. conserved Gly positions). We interpreted our observations and predicted specific structural/functional roles for the respective positions which were then used as the guiding constraints for our coordinate models. 3. Targets with significant fold recognition matches according to publicly available programs/servers: T0044 RTCA (first domain), T0067 PBP, T0077 L30. For this group of targets, manual threading alignments were submitted for fold templates suggested by the UCLA-DOE server [5] and ProCyon [6] methods with high expected accuracy, with the goal to identify clues used by experts aligning manually that could be explored systematically. In particular, we attempted to identify "anchor" positions first, through evolutionary sequence analysis and functional considerations, where possible, and then adjusted the remainder of the sequence-structure alignments based on SURFACE / INTERIOR / ACTIVE SITE assignments [1]. [1] S. A. Benner, G. Cannarozzi, D. Gerloff, M. Turcotte and G. Chelvanayagam (1997).

Bona fide prediction of protein secondary structure using transparent analyses of multiple sequence alignments. Chem. Rev. 97, 2725-2843. [2] G. H. Gonnet, T. F. Jenny, L. J. Knecht, C. Korostensky, M. Hallett. URL: http://cbrg.inf.ethz.ch/ [3] D. L. Gerloff, F. E. Cohen, C. Korostensky, M. Turcotte, G. H. Gonnet and S. A. Benner (1997). A predicted consensus structure for the N-terminal fragment of the heat shock protein hsp90 family. Proteins 27, 450-458 [4] A. Bairoch. URL: http://www.expasy.ch/sprot/prosite.html [5] D. Fischer, D. Rice, D. Eisenberg. URL: http://www.doe-mbi.ucla.edu/frsvr/frsvr.html [6] M. Sippl, H. Floeckner. URL: http://lore.came.sbg.ac.at/People/seb/snoopy.html

The information about sequence conservation, correlation and apolarity is sufficient for recognizing incorrectly-folded protein models.

Osvaldo Olmea and Alfonso Valenci

Protein families are a rich source of information, sequence conservation and sequence correlation are two of the main properties that can be derived from the analysis of multiple sequence alignments. Sequence conservation is related to the direct evolutionary pressure to retain the chemical characteristics of some positions in order to maintain a given function. Sequence correlation is attributed to the small sequence adjustments needed to maintain protein stability against constant mutational drift. It is shown here that conservation and correlation contain sufficient information to detect incorrectly-folded proteins in a significant number of cases. Indeed, they seems to contain as much information as polarity, a property that has been extensively used in this field. The combination of conservation, correlation and polarity leads to an almost perfect discrimination of incorrectly-folded protein models.

Prediction of the structural interaction between different components of the DnaK system and structural interpretation of the functional cycle.

Prediction of the structural interaction between different components of the DnaK system and structural interpretation of the functional cycle. Florencio Pazos and Alfonso Valencia

The Hsc70 chaperon system is essential for protein re-folding and transport in cells. The action of Hsc70 is well regulated by the presence of different effector proteins (DnaJ and GrpE in bacteria) and co-factors (ATP, Mg++ and the refolded peptide). A considerable body of experimental evidence about the structure and function of these proteins has been accumulated during the last years, including the 3D structure of the main components. Still we are far away of fully understanding the mechanism of action and the structural and functional relations between components and states. With the aid of methods for molecular docking based on protein structures and different sources of sequence information (tree-determinants and correlated mutations), we propose a physical model for the interaction of the main components of the system. The model explain part of the functional cycle and leads to detailed predictions that can be further tested experimentally.

Simplified Flexible Geometry Model for Protein Folding

D.J. Osguthorpe

The Reduced Representation Model and Force Field Simplified Geometry Model The model involves representing the backbone of each residue by one sphere, or 'atom', and the sidechains by up to 3 'atoms'. The side chains of Ala, Val, Ile, Ser, Thr and Pro are represented by 1 sphere, Leu, His, Asp, Glu, Asn, Gln, Cys and Met by two spheres and Phe, Tyr, Trp, Lys and Arg by three spheres. The different number of spheres reflects the anisotro- pic nature of the average shape of the corresponding side chains. It also enables assigning different characteristics to parts of the side chain of a residue, for example, the side chain of Arg includes a hydrophobic chain and a polar/charged end. Although in this representation many residues have the same number of atoms, they do not lose their unique identity since they have different parameters. Simplified Potentials The potentials required can be split into three major groups, the virtual internal potentials which stabilise the geometry of the protein, secondary structure stabilisation potentials and the global potentials, which deal with the effects of the environment but do not require the environment to be modelled explicitly. The potential energy function for the model is defined as: E total = E Internal + E Secondary Structure + E van der Waals + E Global Internal Potentials The values of the parameters were derived by fitting observed distributions of the corresponding internals in experi- mental structures and by emulating the energy surface calculated using a full atom model. The internal energy is defined in terms of virtual bond, angles and torsions (or out of plane). A number of functional forms are used, the standard full-atom model harmonic terms, quadratic functions and gaussian functions plus combinations of these terms. Additionally an out of plane-virtual valence angle cross-term is defined. E Internal = E V. bond + E V. angle + E V. torsion + E V. oop + E V. oop X V. angle Virtual angle - virtual angle - virtual torsion angle (theta-theta-phi) cross-terms are defined for dealing with correlations between the two internal valence angles of a torsion angle in the backbone. These are particularly important for turn conforma- tions. Secondary structure energy/Backbone Hydrogen bonding Potentials - 2 - With the simplified geometry model only C alpha atoms exist for the backbone and yet backbone hydrogen bonding is very important in the stabilisation of the standard secondary structures. How- ever, the standard secondary structures have a fixed and specific set of distances between the C alpha atoms. Hence the basic approach was to determine the equilibrium distances between C alpha atoms in 3-10, alpha-helices and parallel and anti-parallel beta-sheets and to use gaussian functions to stabilise these dis- tances. E Secondary Structure = E Helix + E Sheet For the beta-sheets it was also necessary to include some vector terms as well to ensure only when the two strands were aligned was the potential strong. Further improvements were necessary to the sheet potentials after trial folding runs as it became clear that additions were needed to remove conformations created that are never seen in real proteins. It should be noted that in all cases the secondary structure potentials merely stablise distances that are found, this is not a pre-imposition of secondary structure. Indeed the beta-sheet potentials do a full search of all residue pairs to find any that are close enough to form sheets in each energy calculation. Secondary structure prediction energy This is a new term added since CASP2 to account for the observa- tion that certain residues prefer a particular secondary struc- ture which is not accounted for by simple side chain interac- tions, e.g. the significant preference of Ala for helix (possi- bly caused by side-chain entropy). Ala, Lys, Arg, Glu, Gln, Leu and Met are assigned a helix preference, while Val, Ile, Thr, His, Phe, Tyr, Trp and Cys are assigned a strand preference. An overall preference for any residue has been added by stabilising virtual torsions and angles using i-i+2, i+1-i+3, and i-i+3 dis- tances and gaussian functions for both the helical and strand conformations. As individual residue conformations only affect the virtual valence angle, the overall preference is specifically increased only for contiguous pairs of residues which both prefer the helical conformation or both prefer the strand conformation. That is, the two central C alphas of a backbone virtual torsion must both prefer the helical or strand conformation to increase the secondary structure prediction potential of the virtual tor- sion. E Secondary Structure Prediction = E Turn + E Strand Global/Solvation Potentials The remaining potentials are used to represent the non-bonded interactions of the residues with each other and the interactions with solvent. The fundamental idea behind the solvation poten- tials was to use fast approximations to the physical forces involved in real protein structures. Also, as molecular dynamics was seen as one of the primary tools to be used in the - 3 - parameterisation procedure and for first attempts at protein folding the potentials had to have analytical derivatives for speed. Physical Model Solvation Potentials. In this potential model the physical forces of solvation were included using simple potential models. The main idea was that most protein atoms should not have an attractive interaction with other protein atoms, reflecting the fact that the real interac- tions with protein atoms would be replaced by solvent interac- tions if the atom became exposed, hence its overall energy would not change depending on whether it was buried or exposed. How- ever, the atoms should still have excluded volume so a repulsion potential is required at short distances. Just removing the dispersion term from the Lennard-Jones is not good enough as the repulsion component is well above 0 at the minimum of the origi- nal Lennard-Jones potential. What is required is a potential which is zero at the Lennard-Jones minimum and which increases to fit the original Lennard-Jones repulsion as the distance between atoms decreases. In order to achieve this a very high power repulsion is required using the same form as the Lennard-Jones repulsion component, currently a 30th power is being used. Addi- tionally an offset is subtracted from the distance between atoms to increase the repulsion. This potential is used for most atoms, in particular the C alpha backbone atoms and any atom which does not have a specific Lennard-Jones potential. E van der Waals = E Lennard-Jones repulsion only Physical Model Solvation Potentials - Hydrophobicity The next effect to consider is the "hydrophobic" effect. I con- sider this to be associated with two parts, the Van der Waal's potential between atoms (which is attractive) and effects due to the interactions with water. When side-chains are buried in the hydrophobic core of a protein the only interactions available are the standard Van der Waal's interactions, as there is no water present. Hence side-chain atoms of hydrophobic groups were given a standard Lennard-Jones potential with an initial energy assign- ment for interactions between the same atom close to the enthalpy of vapourisation of the most similar hydrocarbon. This would reproduce the energy of the hydrophobic core when hydrophobic side-chains are buried. This determined the potential between the same side-chain atom types. Interactions between dissimilar side-chain atom types were then considered. Analysis of the distribution of side-chain atoms around an atom in known protein structures showed to a first approximation little difference in preference between the atoms. This distribution is not that which is created by rules such as the geometric mean rules. A function was created which would give such as a distribution and this was used to generate the mixed terms for the Lennard-Jones parameters of hydrophobic side chain atoms. - 4 - Having accounted for the potential of hydrophobic side-chains when buried and away from water, a potential for hydrophobic atoms when exposed to water is required. It is only this term which I consider to be truly the "hydrophobic" part, in the sense it reflects the effect of hydrophobic groups on water structure. This was done by introducing a hydrophobic sigmoid potential. (The initial folding work of Levitt had used a sigmoid potential for hydrophobic residues.) In this type of potential the dis- tances from one atom to all other atoms of the same type are com- puted and converted through some form of sigmoid function before being summed to give the potential value for the atom. A final adjustment to the "hydrophobicity" potential was to give certain groups in residues not normally considered hydrophobic a non-zero Lennard-Jones function so that an interaction existed between them and hydrophobic groups. These groups were not included in the sigmoid potential. Such groups were the Ala C beta, the Thr C beta (because of the methyl group), the C beta of the charged amino-acids Asp, Glu, Lys and Arg and Asn and Gln. It also included the C gamma atom of Lys and Arg. Observations of experimental structures and surface accessibility calculations show that these groups are as buried as any of the atoms in the classic hydrophobic side-chains. E Global = E van der Waals + E hydrophobic sigmoid Physical Model Solvation Potentials - Electrostatics The Kirkwood-Tanford model assumes a spherical "protein" which has a low dielectric constant in an environment with a high dielectric constant plus ionic effects, in which point charges are embedded and the paper gives full formulae for the calcula- tion of the electrostatic energy, including self energy. Using model calculations based on these formulae if the charges are exactly on the boundary between the high and low dielectric region, the coulombic interactions between atoms to a very good approximation are given by the standard formula with the dielec- tric constant of the high dielectric region. Hence the first approximation used is to ignore buried salt bridges and consider all charged atomic groups of charged residues to be on the "sur- face" of the protein and hence the simple coulomb law could be used. To take into account of ionic strength effects, which are assumed to have an affect at large distances between charges but not at short range (as the Debye-Huckel theory on which this aspect is based assumes an averaged ionic atmosphere around each charge which is certainly not true for charges on the surface of a protein), a combined coulomb law based on two distance powers and dielectrics was used, the standard distance (power one) and a distance cubed term. The other feature of electrostatics that needs to be covered is the difficulty of burying charges. It is actually a much stronger rule of proteins that the charged group of charged resi- dues is exposed than that the sidechains of hydrophobic residues are buried. Charged groups are only buried if in a salt bridge or extensively hydrogen bonded. The simple electrostatic - 5 - explanation for this is the self-energy of a charge which says it requires a lot of energy to move a charge from a high dielectric region into a low dielectric region. An inverse sigmoid function based on distances between all charge centres of like charge type and residue type. This poten- tial gets larger as atoms get closer together, hence preventing charged groups from being buried. As there is a big difference in surface accessibility between the 4 charged residues, Lys, Glu, Asp, and Arg independent potentials are used for Lys, Asp/Glu and Arg. The Lysine charged end point is the most sol- vent exposed group of proteins, with an average relative surface accessible area of greater than 50%. Glu is next followed by Asp, both in the 45% region, and Arg is the least exposed at around 35%. This is what you would expect based on charge den- sity considerations, the self energy being much greater for a charge field which small and highly charged. The amine group of Lysine is the smallest charged group, with only one heavy atom, the carboxyl spreads the charge further while the guanidinium group charge is spread over a very large area (four heavy atoms). To create a dielectric boundary this inverse sigmoid function is divided by a fixed radius and taken to an arbitrary power. Note that at the fixed radius the value of these terms is one. E Global = E Electrostatic + E inverse charge sigmoid Physical Model Solvation Potentials - Scaling In the low dielectric environment of the folded protein the sta- bility of the backbone-backbone hydrogen bonds is significantly enhanced as these hydrogen bonds are excluded from solvent and a hydrogen bond is essentially an electrostatic interaction. In the unfolded protein the stability of backbone-backbone hydrogen bonds is likely to be similar to that of backbone-water hydrogen bonds, hence there should be no energy stabilising backbone hydrogen bonds. This effect has been included by scaling the backbone hydrogen bond energy term (E Helix and E Sheet) accord- ing to some radius of gyration of the protein, on the assumption that a large radius of gyration indicates an unfolded protein. The radius of hydrophobic gyration was used to measure how far the protein had created a low-dielectric hydrophobic core. The same scaling factor was also used to scale the "self-energy" term, the inverse charge sigmoid, as again charges would only not like to be buried when the hydrophobic core of the protein had been formed. Folding Simulations - Simulated Annealing procedure The starting conformation was an all-extended structure using a rigid geometry procedure based on a standard geometry for the RR model. A random Maxwell-Boltzmann distribution was used - 6 - to assign initial velocities. The initial temperature was set such the average temperature initially was around 340-350K. 84000x5 steps were run before starting cooling. The annealing protocol was first to reduce the total energy by the energy equivalent to 25 degrees of temperature in 84000 steps followed by 84000x4 steps at constant total energy. This was repeated three times. Then the energy was reduced by 12.5K in 84000 fol- lowed by 84000x4 steps at constant total energy, repeated 15 times. Final annealing was by 12.5K in 84000 steps repeated 10 times.

Toward Automatic Methods for Identification of Structural Similarities

Michael B. Bass and Roland L?hy

The process of rapid structural classification of unknown coding sequences is important for identifying novel sequences which might be candidates for development as therapeutics. We describe here a protein threading method using statistical potentials to classify sequences into structural families by comparison to a unique subset of PDB structural database. AmgenThread is a sequence-structure comparison program based on the Needleman-Wunsch algorithm. Global alignments without end penalties are used. The comparisons to known structures are made through the use of three statistical (log-odds) potentials based on surface exposure, pairwise contacts, and Phi-Psi dihedral angle propensities. All statistical potentials are derived from a collection of PDB structures which share less than 35% sequence identity (Hobohm and Sander). The surface exposure term is calculated from the relative surface exposure as compared to Gly-X-Gly using the method of Ponder and Richards. Pairwise contacts are determined by the closest sidechain atom distance between two residues which are separated by at least 5 residues in the primary sequence. The distances are collected into 0.5 ?bins for the statistical sampling. The dihedral angle term was determined using 18 ㅀ bins in each angle. In the comparison to the structural database, a set of unique protein structures sharing less than 45% sequence identity was used (Hobohm and Sander). A second comparison database composed of domains derived from this structural database was also used. Alignment scores were normalized by dividing by the length of the longer of the two sequences. If the Z-score was greater than 5, a successful prediction was reported. If the Z-score was less than 5, a further examination was done to see if the target sequence could be divided into domains which would yield an acceptable Z-score.

Multiple features approach to protein fold recognition.

Krzysztof A. Olszewski

Secondary structure enhanced sequence similarity scoring based fold recognition [Fischer and Eisenberg, 1996] has been demonstrated to be very effective in CASP2 experiment either in combination with other methods [Rice et al., 1997] or alone (in post CASP2 experiment) [Olszewski et al., 1998]. The approach was proven to be successful on the verge of the twilight zone of protein similarity but become more difficult when the homology between the target and the available fold dropped below 12% (as measured by the average percent identity). Interestingly, very often the correct answer is obscured by a few false positive hits, appearing because the scoring system is not selective enough. The possible way to deal with this problem is to use different scoring functions with the goal to discern false positives. In this experiment, the novel approach has been proposed based on rescoring high scoring hits obtained with primary scoring function, with secondary scores derived from the alignment features descriptors. In this approach sequence similarity score enhanced with predicted secondary structure is used to obtain initial order and alignments of putative hits. Then for alignments obtained in the previous step, their characteristic features like gapped and gapless alignment length coverage, gapped and gapless percent identity, mean gapples score per residue, etc. are calculated. It has been shown [Olszewski, in prep.] that such features can be defined as intensive properties (i.e. they do not depend on the length of the alignment), which facilitates statistical analysis of each property with respect to the random alignment model. It also allows to directly compare alignments with drastically different alignment length. For each property, its z-score is calculated, using two models: all alignments model and equal alignments model. E.g. average percent identity relative to the gapples alignment length has been found to be very sensitive for the remote homology in the twilight zone. This property must, however, be accompanied by the high gapples alignment coverage of the hit to be significant. Such concurrent analysis of alignment scores and z-scores is conducted by constructing a series of scattered plots in the application that offer a possibility to track down the location of inherited annotations. The analysis outlined above has been performed for all CASP3 targets with a nontrivial sequence homology to sequences with known structures and predictions for most of them have been submitted. Fischer D. and Eisenberg D., Prot. Sci., 5, 947 (1996) Olszewski K.A., Edwards, D. and Yan L., Theor. Chem. Acc. in press (1998) Rice D.W., Fischer D., Weiss R. and Eisenberg D., Proteins, Suppl.1, 113 (1997)

Neural Network Assignment of Protein Secondary Structure with Increased Predictability

Claus A. Andersen, Soeren Brunak

A large data set of 707 non-homologous protein chains has been analyzed, mainly with respect to the mapping between secondary structure and amino acid sequence. It is shown that the commonly used assignment scheme DSSP performs a sub-optimal capping assignment with respect to the sequence signals. In contrast to other schemes DSSP has an excess of length four $\alpha$-helix assignments. This is shown not to have a structural basis with regard to the C$_{\alpha}$-coordinates or the $\phi$/$\psi\mbox{-angles}$. The main factors to achieve a high secondary structure prediction performance from the amino acid sequence have been identified as the use of evolutionary information in the form of sequence profiles and the selection of predictable categories. Evaluating predictors from their total percentage performance is argued to be sub-optimal, since this measure favors over-prediction of the largest structural class \textit{i.e.}\ coil. In order to improve secondary structure prediction a novel approach is put forward, where new assignments are evaluated by their predictability from the amino acid sequence. These new assignment algorithms are constructed by training neural networks to predict the DSSP assignments from the 3-dimensional structural data. Two representations of the structural data have been tested as inputs to the neural network, which has shown that an assignment based on the C$_{\alpha}$-coordinates is more predictable than one based on the $\phi$/$\psi\mbox{-angles}$. By evaluating the neural network assignments on their predictability the best network configuration is selected. This new assignment is more predictable than that of DSSP tested on the 707 protein chain data set. The prediction performance increases to $Q_3^{tot}=\,$68\%, $C_{\alpha}=\,$0.55, $C_{\beta}=\,$0.42, and $C_{coil}=\,$0.49 in a ten-fold cross-validation, without the use of evolutionary sequence profiles or structure-to-structure postprocessing. Through an inspection of a neural network without hidden neurons and its assignment, it is found that 85\% of DSSP's $\alpha\mbox{-helix}$, $\beta\mbox{-sheet}$ and coil classifications can be learnt by a network performing a linear separation of the C$_{\alpha}$-coordinates from small windows of local sequence segments. A test of the neural network assignment's predictability, on the 126 protein Rost and Sander data set, using evolutionary sequence profiles

and a simple multi-level setup, increases the correlation coefficients with respect to other predictors: $Q_3^{tot}=73.5\%$, $C_{\alpha}=0.66$, $C_{\beta}=0.51$, and $C_{coil}=0.52$ in a seven-fold cross-validation. The improvement is mainly due to the exclusion of 3$_{10}\mbox{-helices}$ from the helix class, but further improvements are outlined by using several sets of neural network assignments.

Neural Network Assignment of Protein Secondary Structure with Increased Predictability

Claus A. Andersen and Soeren Brunak

A large data set of 707 non-homologous protein chains has been analyzed, mainly with respect to the mapping between secondary structure and amino acid sequence. It is shown that the commonly used assignment scheme DSSP performs a sub-optimal capping assignment with respect to the sequence signals. In contrast to other schemes DSSP has an excess of length four α-helix assignments. This is shown not to have a structural basis with regard to the C$_{\alpha}$-coordinates or the ϕ/$\psi\mbox{-angles}$. The main factors to achieve a high secondary structure prediction performance from the amino acid sequence have been identified as the use of evolutionary information in the form of sequence profiles and the selection of predictable categories. Evaluating predictors from their total percentage performance is argued to be sub-optimal, since this measure favors over-prediction of the largest structural class \textit{i.e.}\ coil. In order to improve secondary structure prediction a novel approach is put forward, where new assignments are evaluated by their predictability from the amino acid sequence. These new assignment algorithms are constructed by training neural networks to predict the DSSP assignments from the 3-dimensional structural data. Two representations of the structural data have been tested as inputs to the neural network, which has shown that an assignment based on the C$_{\alpha}$-coordinates is more predictable than one based on the ϕ/$\psi\mbox{-angles}$. By evaluating the neural network assignments on their predictability the best network configuration is selected. This new assignment is more predictable than that of DSSP tested on the 707 protein chain data set. The prediction performance increases to $Q_3^{tot}=\,$68\%, $C_{\alpha}=\,$0.55, $C_{\beta}=\,$0.42, and $C_{coil}=\,$0.49 in a ten-fold cross-validation, without the use of evolutionary sequence profiles or structure-to-structure postprocessing. Through an inspection of a neural network without hidden neurons and its assignment, it is found that 85\% of DSSP's $\alpha\mbox{-helix}$, $\beta\mbox{-sheet}$ and coil classifications can be learnt by a network performing a linear separation of the C$_{\alpha}$-coordinates from small windows of local sequence segments. A test of the neural network assignment's predictability, on the 126 protein Rost and Sander data set, using evolutionary sequence profiles and a simple multi-level setup, increases the correlation coefficients with respect to other predictors: $Q_3^{tot}=73.5\%$, $C_{\alpha}=0.66$, $C_{\beta}=0.51$, and $C_{coil}=0.52$ in a seven-fold cross-validation. The improvement is mainly due to the exclusion of 3$_{10}\mbox{-helices}$ from the helix class, but further improvements are outlined by using several sets of neural network assignments.

Homology modeling for target 47

zhen wang

METHOD Comparative modeling was used to predict the structure of T0047. A METHOD database search identified that t0047 belongs to the lipocalin family. METHOD All the proteins in this family have eight stranded antiparallel METHOD beta_barrel structures. They bind small hydrophobic molecules. METHOD 1mup was selected as the template. T0047 has 64% identities and 78% METHOD similarity with 1mup. METHOD The method of mainchain building: Because of the large number of METHOD identities, the mainchain conformation of 1mup was used throughout. METHOD Method of sidechain building: Minimum perturbation (MP) method METHOD implemented by the program MUTATE was used to generate initial METHOD sidechains for all the residues. The MP method preserved the METHOD same Chi angles for the t0047 from 1mup. METHOD Graph method (Samudrala and Moult, JMB, 1998, 279, METHOD 287-302) was then used to refine the sidechain conformations. METHOD Computing limitations made it necessary to treat no more than METHOD about 12 residues at a time. Therefore, the structure was METHOD divided into 20 groups, each group containing 6 to 12 residues. METHOD The sidechain conformations in each group were optimized by the METHOD graph method against the background of the current model. Up to METHOD 6 sidechain conformations were selected per residue (Samudrala METHOD and Moult, Protein Engineering, in press). The 100 best scoring METHOD cliques were selected and reevaluated with the knowledge-based METHOD potential (Samudrala and Moult, JMB, 1998, 275, 895-916). The METHOD conformation from the best scoring clique was then incorporated METHOD into the model. METHOD Use was also made with the published picture of the ligand binding METHOD cavity (Bocskei and Colin, etc. Nature, 1992, 360(12):186-189). METHOD One sidechain (residue 60) was adjusted to fit the picture, rather METHOD than accepting the generated conformation. METHOD The model was checked by using PROCHECK, cavity inspection METHOD using QUANTA and a check for large unfavorable electrostatic METHOD interactions (Toner-Oliva and Moult, submitted). Sidechains with METHOD a large amount of exposure hydrophobic areas were also inspected. METHOD As a result of these checks, a few sidechains were adjusted by METHOD hand.

All-Atom Threading Models

Levitt,Samudrala with Xia,Brenner

The method used for the submissions by the Levitt group (7936-3040-4355) for the fourteen targets, T0048, T0055, T0057, T0062, T0064, T0065, T0068, T0069, T0070, T0074, T0076, T0079, T0082, and T0085, is as follows: (1) Find a putative template sequence of known structure for each query sequence by a combination of three methods: (a) A FASTA search against the sequence file from the currect set of PDB coordinates (ftp://pdb.pdb.bnl.gov/pub/pdb_seqres.txt). (b) A PSI-BLAST search invoving a database of 267000 non-redundant protein sequences to make profiles, followed by a search of these templates against the pdb sequences. (c) An intermediate sequence FASTA search using a sequence data base of 250,000 non-redundant protein sequences and the scop 1.37 super- family of possible templates. In general, the most useful method was the simplest, a FASTA search against the latest sequences from all the pdb files. (2) Choose a template PDB file. This was done by aligning all members of the scop superfamily that contained the putative template match. This multiple sequence alignment was done using Gotoh's PRRP method (Gotoh, 1996). We used the Blosum50 matrix but otherwise all parameters were the default values. (3) Add the observed secondary structure for the structured sequences and the predicted secondary structure for the query to the multiple alignment. Use Eric Sonnhammer's Belvu program (Sonnhammer 1998) to view the alignment. Delete unwanted alignments. Choose a final template sequence and the alignment of the query to it. In a few cases this alignment was hand-edited to ensure that deletion occured across loops (4) Use a Monte Carlo perturbation algorithm to generate approximately 1000 different alignments. Use SegMod (Levitt, 1992) and Encad (Levitt et al. 1996) to build all atom models for each of these alignment. In a few cases, models were built from different templates. (5) Rank these models using the Samudrala & Moult (1998) scoring function. The submitted entries included models built from the initial un- perturbed alignment as well as four models from perturbed alignments with the most favorable scores. In all case, models built from the perturbed alignments had better scores than the models built from the initial alignment. Gotoh O. Significant Improvement In Accuracy Of Multiple Protein Sequence Alignments By Iterative Refinement As Assessed By Reference To Structural Alignments. J Mol Biol. 264:823-838 (1996). Levitt, M. Accurate Modelling of Protein Conformation by Automatic Segment Matching. J. Mol. Biol. 226, 507-533 (1992). Levitt, M., M. Hirshberg, R. Sharon and V. Daggett. Potential Energy Function and Parameters for Simulations of the Molecular Dynamics of Proteins and Nucleic Acids in Solution. Computer Physics Communications, 91, 215-231 (1995). Samudrala, R, and J. Moult. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J. Mol. Biol. 275, 893-914 (1998). Erik Sonnhammer. Belvu. http://www.sanger.ac.uk/Software/Pfam/help/belvu_setup.shtml (1998).

Prediction of secondary protein structure and fold recognition using nearest-neighbor approach

Victor Solovyev and Asaf Salamov

To predict secondary structure of the presented targets we have used our new nearest-neighbor approach (SSPAL) shortly described in (Salamov,Solovyev,1997). This approach is different from the NNSSP method (Salamov,Solovyev,1995), which was among the top four methods of secondary structure prediction on the previous Asilomar meeting (Lesk,1998). To select neighbors the new approach uses local alignments instead of fixed length segments. For scoring an aligned position of the query sequence with the position of database sequence the hybrid scoring system was used. It combines the local environment score (Bovie et al,1991; Salamov, Solovyev,1995) with the score provided by sequence similarity matrix of Gonnet and colleagues (1992). We compute 50 best non-intersecting local alignments of the query sequence with each sequence from a set of proteins with known 3D structures. Each position of the query sequence is aligned with the database amino acids in a-helical, b-strand or coil states. The prediction type of secondary structure is selected as the type of aligned position with the maximal total score. Additionally the Blastp (Altschul et al.,1997) search in OWL nonredundant protein sequence database (Bleasby and Wootton, 1990) was applied to get multiple alignment of proteins similar with the target sequence. The multiple sequence alignment has been used as input of the method for secondary structure prediction. Prediction of secondary structure by the SSPAL method is available via Baylor College of Medicine and The Sanger Centre Computational Genomic

Group Web servers (http://dot.imgen.bcm.tmc.edu:9331/pssprediction/pssp.html or http://genomic.sanger.ac.uk/). Using combination of three potentials including local environment score, similaruty of secondary structure predicted by SSPAL and the secondary structure of database protein and mutation matrix score we predict fold selecting the protein with the best combined alignment score from the database of about 1000 known folds. We have presented to CASP3 secondary structure prediction for all given targets as well as fold prediction for all target except the last seven (according the expiration date).

References: Altschul S. F. et al.(1997) Nucl. Acids Res., 25: 3389 - 3402. Bowie, J.U., Luthy, R., Eisenberg, D. (1991) Science 253:164- 170. Bleasby, A.J. and Wootton, J.C. (1990) Prot. Engng. 3(3):153-9. Gonnet, G.H., Cohen, M.A., Benner, S.A. (1992) Science 256: 1433-1445. Lesk A. (1987) Proteins, Suppl.1, 151-166. Salamov A., Solovyev V.(1997) J.Mol. Biol., 268, n 1, 1997, 31-36. Salamov A., Solovyev V.(1995) J.Mol. Biol., 247, n 1, 1997, 11-15.

CASP4: CACASP or CAFASP?

daniel fischer

CASP4: CACASP or CAFASP? As a complement to the enormous value of this event, here I propose to discuss the differences between the Critical Assessment of COMPUTER AIDED Structure Prediction (CACASP) and the Critical Assessment of FULLY AUTOMATED Structure Prediction (CAFASP). In the current recipee of CASP, predictors submit their predictions using whatever techniques they choose; some report the exact results of fully automated methods without any changes; some "edit" the results from the programs and create a "computer aided" + human prediction; and some use mostly their brains with little aid from computers. Thus, the current recipee is assessing the performance of humans and does not allow for the strict assessment of the methods themselves. The latter is what ultimately non-expert users are interested in. Predictions in which human intervention has played a role can not always be objectively assessed, and in most cases, they are hard, if not impossible to reproduce by others. The value of the current critical assessment is thus not fully exploited. What I propose is to take advantage of the invaluable framework provided by CASP4 and run in parallel a CAFASP evaluation. To this end, for each target in the fold recognition track, results of fully automated programs and servers need to be compiled and made available thru the web to all, and when the structures are released, an assessment of the methods used will be carried out and published in the web. This would make the "cacasp" section more interesting, because the automatic results will be available to all, and individuals using this information coupled with their newer, not yet available methods will be able to demonstrate their judgement beyond what can be obtained with current automatic methods. A similar "cafasp" section has already been done in casp3 for secondary structure prediction. In addition to the value of assessing fully automated methods, cafasp may encourage the developers of fold recognition methods to automate them and made them available to the community, which ultimately is the goal of science. Programs and servers willing to participate in cafasp4 are invited to help the planning.

Are predicted structures good enough to preserve functional sites?

Liping Wei, Enoch S. Huang, & Russ B. Altman

The quality of predicted structures are currently gauged primarily with numerical metrics that judge the quality of alignments and root-mean-squared deviation (RMSD) relative to the native structure. However, a principal goal of structure prediction is the elucidation of function. We have studied the ability of computed models to preserve the microenvironments of functional sites. In particular, we analyzed 653 model structures (generated using an ab initio folding protocol) of a calcium binding protein, and assessed the degree to which calcium binding sites were recognizable. We found that, while some model structures preserve the calcium binding microenvironments, many others, including some with low RMSD, do not. There is very weak correlation between the overall RMSD of a structure and the preservation of calcium binding sites. Only when the quality of the model structure is high (local RMSD less than 2 Angstroms) does the modeling of the binding sites become reliable. We conclude that protein structure prediction methods need to be assessed in terms of their preservation of functional sites. For detailed analysis such as the precise localization of functional sites, which is important in protein design, high-resolution structures at the accuracy of 1-2 ?are required. The correct modeling of loop regions and side chain atoms is critical.

Protein modeling by homology augmented by ab initio and energy-based methods

G. Raghunathan, M.D. Shenderovich, M. Prabhakaran, M.J. Dudek, C.L. Fisher, J. Zheng, J. Kottalam, B. Vessal and K. Ramnarayan

We have used a combination of both heuristic methods that rely on our knowledge of protein structure and folding, as well as objective and automated criteria in order to build and evaluate structures. We predicted structures for 6 entries in CASP3. Experimental results for one of the structures has been withheld by organizers. RMS deviations between predicted and experimental backbone coordinates for 4 out of 5 predicted structures are 1.4, 1.5, 1.9 and 2.6 A. Sequence alignment and secondary structure prediction are used to select protein family and fold to which the given sequence might belong. Functionally important residues are aligned using multiple sequence alignment methods. Models consistent with alignment are constructed. A suite of quality control methods are used to check for, among other things, packing, solvent accessibility, local strain in structures and alternate placement of loops using a loop-building procedure. Stereochemical permissibility of structures is evaluated using the PROCHECK program, and energy-based methods are used to improve their quality. A comparison of the experimental and predicted structures is discussed along with specific modeling, refinement and quality control aspects that we use in protein modeling.

Selecting sequence to structure alignment by model building

Ceslovas Venclovas, Krzysztof Ginalski and Krzysztof Fidelis

Molecular models for four target proteins (T0047, T0048, T0055 and T0070) were built using related structures as templates. These targets and their closest by sequence structural relatives had sequence identity of 62%, 27%, 24% and 19% respectively. The modeling method was based on the idea that the correctness of sequence to structure alignment is critical to the quality of the model while other steps such as loop building and/or energy minimization can at present contribute little if anything. Multiple sequence alignments were used to identify regions where alternative alignment choices were possible. Variants suggested by aligning multiple sequences with a number of different gap opening/extension penalties were used to build test models. For apparently structurally conserved regions where no reliable automatic alignment could be generated, a combination of secondary structure prediction and manual alignment was used. The final sequence to structure alignment was selected after testing all relevant alignments by building 3D models and evaluating their consistency with the parent structure. Models were built using Homology module of InsightII (MSI Inc.). For structurally conserved regions backbone conformation was preserved and only the side chains were substituted. Modeling of loops was skipped for most models built to test a particular alignment. Loop conformations for the final models were selected from database of protein structures. After the coordinates were assigned to the target sequence, rotamers were rebuilt using backbone dependent rotamer library [1]. To preserve conserved contacts some rotamers were set back manually. Besides visual inspection, structural consistency of the models was evaluated both for general fitness as well as for structural details using mainly ProsaII [2] and Whatcheck [3]. Final models were subjected to 100 or less steps of steepest descent minimization using InsightII Discover to remove remaining steric clashes. References: 1. Bower M.J., Cohen F.E., Dunbrack R.L. Jr. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. J.M.B., 1997, 267(5):1268-1282. 2. Sippl M.J. Recognition of errors in 3-dimensional structures of proteins. PROTEINS, 1993, 17(4):355-362. 3. Hooft R.W., Vriend G., Sander C; Abola E.E. Errors in protein structures. Nature, 1996, 381(6580):272.

Protein Structure Prediction using a Combination of Sequence and Structure Similarity Search Techniques

Kristin K Koretke, Richard Copley, Robert Russell, and Andrei N Lupas

Each prediction was started with a PSI-BLAST search against the nonredundant database to identify homologous sequences. The default parameters, including the BLOSSUM62 similarity matrix with gap penalties of 11 to open a gap and 1 to extend and an E-value threshold of 0.001 for inclusion in successive iterations, were used for every run. Sequences with an E value of 1.0e-03 or less were imported into MACAW A multiple sequence alignment was constructed within MACAW using pairwise segment overlap and Gibbs sampling with the BLOSSUM62 similarity matrix. This alignment was used to create an HMM profile, and HMMer was used to search the PDB database for distant homologs. The multiple sequence alignment was also used to obtain a secondary structure prediction. In instances where our alignments differed from the default alignment found on the JPRED server, a JPRED prediction was performed, otherwise we simply used that provided Once we had both a multiple sequence alignment and a secondary structure prediction, we examined the data for any notable sequence or structural motifs. We then explored the output from our PSI-BLAST and HMMer searches to identify any significant hit having a crystal structure. For the cases in which the PSI-BLAST and/or HMMer searches did not generate a possible fold within the significant hits, we probed the trailing ends of the PSI-BLAST output (E-values between 0.1 and 10.0) for sequences with very distant homology but conservation of important motifs. If a sequence looked promising, it was used as a query sequence for a PSI-BLAST search against the non-redundant database. A similar protocol of generating a multiple sequence alignment, HMM, and secondary structure prediction was applied to the results of this search. If we could not find any promising leads within the trailing ends of our PSI-BLAST searches, we examined the multiple sequence alignments for significant patterns. Any promising patterns were used as input to the GCG Findpatterns program. If this search method did not yield a hit to a sequence with a known function, we submitted the secondary structure prediction from JPRED as input to the MAP program. If we could not resolve an answer from the MAP output, we conducted visual inspections of the SCOP database looking for folds that would fit the secondary structure prediction. If these searches yielded nothing, we predicted 'novel fold'. We used biochemical information identified through MEDLINE searches where available. The target proteins predicted using sequence search methods were aligned to the template with MACAW (and thus aided by the multiple sequence alignments that had been generated previously) or HMMer. The target proteins predicted using MAP were aligned to the template initially either according to an alignment from HMMer, or by using the default alignment from MAP. The proteins predicted by visual inspection of the SCOP database were aligned to the template by overlaying of conserved hydrophobic residues. All alignments were adjusted manually around the gap regions based on their location within the scaffold, and additionally if automated alignments appeared to be incorrect.