Critical Assessment of Techniques for Protein Structure Prediction

Seventh Meeting

ASILOMAR CONFERENCE CENTER
PACIFIC GROVE, CALIFORNIA
NOVEMBER 26-30, 2006

Abstracts

3Dpro - 500 models for 100 3D targets
FOLDpro - 600 models for 100 3D/100 DP targets
ABIpro (server, 3D) - 495 models for 99 3D targets

3D Structure Prediction Using FOLDpro, 3Dpro, and ABIpro

Jianlin Cheng, Arlo Randall, Mike Sweredoski and Pierre Baldi
Institute for Genomics and Bioinformatics, School of Information and Computer Science
University of California Irvine, CA 92697

Three servers (FOLDpro, 3Dpro, and ABIpro) from our group participated in 3D structure prediction in CASP7. FOLDpro is a template-based method using a machine learning approach to rank templates [2]. ABIpro is an *ab initio* method. 3Dpro is a combination of the template-based method and the *ab initio* method. Here we briefly describe the protocol of each server.

FOLDpro

FOLDpro makes prediction in four steps. First, it extracts pairwise similarity features for a query and all templates in the library using alignment tools and structural feature predictors. It also uses PSI-BLAST [1] to search the query against the template database.

Second, a support vector machine (SVM) integrates pairwise features to evaluate the structural relevance of the query and the templates (in the same fold or not). It uses relevance scores to rank the templates. SVM ranking may not always put the best templates on the top of the positive template list. For instance a template in the same fold as the query may be ranked before a template in the same family. So the positive templates are re-ranked by the e-values of PSI-BLAST search if available.

Third, FOLDpro generates an alignment between the query and each of the top 5 templates respectively. For templates that can be found by PSI-BLAST, PSI-BLAST alignments are used. For harder templates, FOLDpro uses a global profile-profile alignment method COACH [3] to generate the alignments between the query and the templates.

Fourth, FOLDpro uses Modeller [4] to build 3D structure for the query, based on its alignments with the templates. Multiple significant templates are combined to generate structures.

ABIpro

ABIpro is an *ab initio* tertiary structure predictor. The energy function is composed of terms from predicted structural features, physical forces, and statistical analysis of PDB proteins. The conformational search is performed using simulated annealing and a segment library.

The search energy includes terms for the following predicted structural features: secondary structure (SSpro), relative solvent accessibility (ACCpro), and residue level contacts (CMAPpro) [5]. The physical terms include hydrogen bonding, van der Waals interactions, electrostatics, and solvation effects. The potential also includes statistical terms for residue solvent environment, local structure independent residue pairing [6], and local structure dependent residue pairing [7].

The search is performed in two phases of simulated annealing. Both phases use a linear cooling schedule and use the same temperature settings. The first phase uses a zero weight for atomic repulsion. The second phase includes the repulsive terms and scales up the weights for other terms to decrease the move acceptance rate. The main move type is fragment replacement with fragment lengths from three to nine residues [8]. Many models are generated using random seeds and those with the lowest energy are selected.

3Dpro

3Dpro is a combination of template-based method and *ab initio* method. It first uses template-based method (the same as FOLDpro, but run independently on slightly different database) to identify templates. If positive templates (SVM score > 0) are found, it uses the same protocol of FOLDpro to make predictions and *ab initio* method is not used. If no positive templates are found, *ab initio* method (the same as ABIpro) is invoked to generate two *ab initio* models. In this situation, three (or four) template-based models and two (or one) *ab initio* models were submitted to CASP7.

1. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W. and Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res., vol. 25, pp. 3389-3402.
2. Cheng J. and Baldi P. (2006) A Machine Learning Information Retrieval Approach to Protein Fold Recognition. Bioinformatics, vol. 22, no. 12, pp. 1456-1463.
3. Edgar R.C., and Sjölander K. (2004) COACH: profile-profile alignment of protein families using hidden Markov Models. Bioinformatics, vol. 20, pp. 1309-1318.
4. Sali A., and Blundell T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol., vol. 234, pp. 779-815.
5. Cheng J., Randall A., Sweredoski M., and Baldi P. (2005) SCRATCH: a Protein Structure and Structural Feature Prediction Server. Nucleic Acids Research, vol. 33 (web server issue), w72-76.
6. Simons K.T., Ruczinski I., Kooperberg C., Fox B.A., Bystroff C. and Baker D. (1999) Improved recognition of native-like protein structures

using a combination of sequence-dependent and sequence-independent features of proteins. Proteins, vol. 34, pp. 82-95.
7. Zhang Y., Kolinski A. and Skolnick J. (2003) TOUCHSTONE II: A New Approach to Ab Initio Protein Structure Prediction. Biophysical Journal, vol. 85, pp. 1145-1164.
8. Simons K.T., Kooperberg C., Huang E. and Baker D. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol, vol. 268, pp. 209-225.

AMBER/PB - 96 models for 1 3D/ 91QA targets

Quality Assessments of Server Results with AMBER/PBSA

M.J. Hsieh, E. Chanco and R. Luo
University of California Irvine
rluo@uci.edu

We have constructed a model scoring scheme given either alignments or tertiary structures for CASP7. The inputs for our scoring scheme are the predicted protein structures from individual servers. These predictions (tertiary structures, main-chain structures, or alignments only) are used to build all-atom models by MODELLER.[1] The all-atom models are then energy-minimized with SANDER in the AMBER suite[2] before scoring. The AMBER/PBSA scoring function,[3] based on a revised ff99 all-atom AMBER force field[4] and the PBSA solvation model,[5] is used to evaluate free energies of minimized structures with PBSA in the AMBER suite.[2]

The resolution of the scoring scheme is found to be about 0.05 GDT value due to the all-atom reconstruction procedure used. The prediction accuracy of AMBER/PBSA is about 70%, based on tests with continuous targets that are not in the new fold category in both CASP5 and CASP6. This is higher than individual servers (56.4% in CASP5, 45.5% in CASP6). AMBER/PBSA also performs much better than two widely used scoring functions DFIRE and ROSETTA tested under the same condition.

In the QA category of CASP7, only the first two predictions per server are evaluated due to the computational cost of the scoring scheme. All free energies are then translated into P-values based on the extreme value distribution.

1. Martí-Renom M.A., Stuart A.C., Fiser A., Sánchez R., Melo F. & Sali A. (2000) Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct. 29, 291-325.
2. Case D.A., Cheatham T.E., Darden T., Gohlke H., Luo R., Merz K.M. Jr., Onufriev, A., Simmerling, C., Wang, B. & Woods, R. (2005) The Amber biomolecular simulation programs. J. Computat. Chem. 26, 1668-1688.
3. Hsieh M.J. & Luo, R. (2004) Physical scoring function based on AMBER force field and Poisson-Boltzmann implicit solvent for protein structure prediction. Proteins. 56, 475-486.
4. Lwin T.Z., Zhou R., Luo R. (2006) Is Poisson-Boltzmann theory insufficient for protein folding simulations? J. Chem. Phys. 124, 039402.
5. Lu Q. & Luo R. (2003) A Poisson–Boltzmann dynamics method with nonperiodic boundary condition. J. Chem. Phys. 119:21, 11035-11047.

Andante - 552 models for 100 3D/ 42 DP/3 FN/1QA targets

Tertiary Structure Prediction of CASP7 Targets Using Exhaustive Modeling and Evaluation

K. Tomii[1], C. Motono[1], M. Sato[1], M. Ota[2], T. Hirokawa[1], P. Horton[1] and Y. Akiyama[1]
[1] - *Computationa1 Biology Research Center, National Institute of Advanced Industrial Science and Technology, 2-42 Aomi, Koto-ku, Tokyo, Japan*
[2] - *Global Scientific Information and Computing Center, Tokyo Institute of Technology, O-okayama, Meguro-ku, Tokyo 152-8550, Japan*
k-tomii@aist.go.jp

According to the activity of world-wide structural genomics, a huge number of protein structures have been determined and deposited in the Protein Data Bank (PDB). Consequently, the fold recognition approach is becoming more effective. In this method, the template structure for the query is selected from the PDB, the 3D model is built based on their alignment and the model is evaluated.

We prepared the following three steps, as in previous CASPs, to identify proper template(s) and to produce target-template alignment(s). To complete these three processes (semi)automatically, we constructed a prediction pipeline, FORTE-SUITE. First, four FORTE series[1-2], which are systems of profile-profile alignment for protein fold recognition, are performed for each target sequence. We used FORTE1, FORTE2, FORTE1T, and FORTE-H for this purpose. FORTE1 and FORTE2 provided target-template alignments that are publicly available in the server category of CASP7. In addition to those alignments, we sampled more alignments using a newly developed substitution matrix (see below). Then, based on those alignments, we constructed and exhaustively evaluated 3D models with MODELLER[3]. According to the Z-scores calculated using the FORTE series, we separately treated easy and hard

targets. For easy targets (generally Z>=8), 10 models were built for each alignment of top 100 proteins in the FORTE library. The 10 models for each alignment of top 500 proteins were constructed for hard targets. Finally, submission candidates among those models were selected using CQS calculated by Verify3D[4] and Prosa2003[5] programs and the new evaluation function, LIBRA_rotamer[6], which we describe below.

We developed our models in terms of the structural quality scores by sampling more alignments after identification of the proper templates. Sampling more alignments was done using the profiles derived from amino acid sequences with various diversity for both targets and templates proteins, or through human intervention in some cases.

We have improved our methodology, especially for the following four directions.

First, to produce a target-template alignment, we also used a substitution matrix that was specially designed for aligning a pair of distantly related protein sequences. This matrix is useful for improving the alignment accuracy when we align two distantly related proteins (unpublished data). In some targets, the matrix yielded alignments, based on which we can obtain the models with highest scores.

To enhance our exhaustive modeling approach, we constructed a high-throughput method of FORTE-SUITE. We were able to build 10 models automatically for the top 500 (or more) alignments obtained by FORTE series. For building as precise 3D-models as possible, we used multiple templates when we were able to use structural information of the same family or a superfamily in PDB.

We introduced a new evaluation function, LIBRA_rotamer, to improve the process of model selection; LIBRA_rotamer checks sidechain interactions, hydration, local propensities, and repulsions of 3D-models based on the 56 rotamers. More details are described in our abstract of the category for quality assessment of models (team name: largo).

In addition to using this new scoring function, we calculated scores, which were averaged over (usually 10) models based on an alignment, to evaluate 3D-models more precisely. As a guidance of model selection, we used the averaged scores instead of a score for each model, which is effective to enhance prediction accuracy, especially for easy targets, according to our results. Using this new protocol when we tested the effectiveness of our new protocol of model selection with CASP6 targets as a benchmark, we attained 4% better results than when using the previous protocol.

1. Tomii K. & Akiyama Y. (2004) FORTE: a profile-profile comparison tool for protein fold recognition. Bioinformatics 20, 594-595.
2. Tomii K., HirokawaT. & Motono C. (2005) Protein structure prediction using various profile libraries and 3D verification. Proteins 61, 114-121.
3. Marti-Renom M.A., Stuart A., Fiser A., Sánchez R., Melo F. & Sali A. (2000) Comparative protein structure modeling of genes and genomes. Annu. Rev. Biophys. Biomol. Struct. 29, 291-325.
4. Eisenberg D., Luthy R. & Bowie J.U. (1997) VERIFY3D: assessment of protein models with three-dimensional profiles. Methods Enzymol. 277, 396-404.
5. Sippl M.J. (1993) Recognition of errors in three-dimensional structures of proteins. Proteins 17, 355-362.
6. Ota M., Isogai Y. & Nishikawa K. (2001) Knowledge-based potential defined for a rotamer library to design protein sequences. Protein Eng. 14, 557-564.

AMU-Biology - 322 models for 92 3D/ 19 FN targets

Combination of template-based and template-free modeling

J. Czwojdrak, U. Baraniak, K. Kaminska, J.M. Bujnicki and A.M. Czerwoniec

Bioinformatics Laboratory, Institute of Molecular Biology and Biotechnology, Adam Mickiewicz University, Umultowska 89, PL-61-614 Poznan, Poland
anna.czerwoniec@amu.edu.pl

In the seventh Critical Assessment of techniques for protein Structure Prediction (CASP7), the AMU-Biology group used the combination of the 'FRankenstein's Monster' approach for template-based modeling (Kosinski, 2003) with the ROSETTA method for de novo modeling (Simons, 1997) to predict the tertiary structure of full-length targets of all categories.

The first step was to recognize structural homologs and generate target-template alignments using a number of fold-recognition methods via the GeneSilico MetaServer (Kurowski and Bujnicki, 2003; http://www.genesilico.pl/meta/). The target alignments were converted into preliminary models using MODELLER (Fiser and Sali, 2003). The preliminary models were evaluated according to knowledge-based potentials implemented in the COLORADO3D server (Sasin and Bujnicki, 2004) to enable discrimination of fragments that are likely to be erroneous. After superimposing the best models, hybrid models were constructed and used to guide modifications of the original target-template alignments. The refinement of models involved iterative model building, evaluation, and realignment. At this step we also used external information: secondary structure predictions, conservation of fragments and putative catalytic residues, and constraints on the

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

1. Kosinski J., Cymerman I.A., Feder M., Kurowski M.A., Sasin J.M., Bujnicki J.M. (2003) A 'Frankenstein's monster' approach to comparative modeling: merging the finest fragments of fold-recognition models and iterative model refinement aided by 3D structure evaluation. Proteins 53 Suppl 6:369-79.
2. Simons K.T., Kooperberg C., Huang E., Baker D. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol. 268(1):209-25.
3. Kurowski M.A., Bujnicki J.M. (2003) GeneSilico protein structure prediction meta-server. Nucleic Acids Res. 31(13):3305-7.
4. Fiser A., Sali A. (2003) Modeller: generation and refinement of homology-based protein structure models. Methods Enzymol. 374:461-91.
5. Sasin J.M., Bujnicki J.M. (2004). COLORADO3D, a web server for the visual analysis of protein structures. Nucleic Acids Res. 32(Web Server issue):W586-9.

Avbelj - 22 models for 7 3D targets

Predictions of three-dimensional structures of proteins using Monte Carlo simulations and electrostatic screening model

F. Avbelj and T. Urbič
National Institute of Chemistry
Franc.Avbelj@ki.si

Three-dimensional structures of proteins are predicted *ab initio* using torsion space Monte Carlo simulations. The method is based on the electrostatic screening model of backbone conformational preferences (ESM)[1-6]. The energy function in the Monte Carlo procedure contains: main-chain electrostatic interactions, electrostatic solvation free energies of main-chain atoms, and hydrophobic interactions. The electrostatic interactions are calculated using Coulomb's law with a dielectric constant of 1. The electrostatic solvation free energies (ESF) are calculated using the finite difference Poisson-Boltzmann model (DelPhi) with PARSE parameter set[7].

Torsion space Monte Carlo simulations of small proteins are performed using hierarchic condensation. In the first phase of simulation only the local electrostatic energies and backbone solvation free energies of residues are activated. In this phase the β-strands are formed. In the second phase of simulation the main-chain hydrogen bonds are included in the energy function. In this phase α-helices and hairpins are formed. In the third phase of simulation the hydrophobic interactions are included in the energy function. In this phase α-helices and β-strands gradually condense into compact structures.

A number of independent Monte Carlo simulations (~10000) are performed. All heavy atoms and polar hydrogen's are included in the simulations. Only torsion angles are allowed to vary during simulations. Hard sphere repulsion is enforced by discarding conformations with steric clashes. Pairs of atoms related by torsion angles are not checked for steric clashes. Conformational space is sampled by varying torsion angles of proteins using different types of moves. The Metropolis criterion is used to decide whether to accept or reject the move. Temperature was 300 K.

1. Avbelj F. and Moult J. (1995) Role of electrostatic screening in determining protein main-chain conformational preferences. Biochemistry, 34, 755-764.
2. Avbelj F. and Fele L. (1998) Role of main-chain electrostatics, hydrophobic effect, and side-chain conformational entropy in determining the secondary structure of proteins. J. Mol. Biol., 279, 665-684.
3. Avbelj F. (2000) Amino acid conformational preferences and solvation of polar backbone atoms in peptides and proteins. J. Mol. Biol., 300, 1337-1361.
4. Avbelj F. and Fele L. (1998) Prediction of the three dimensional structure of proteins using the electrostatic screening model and hierarchic condensation. Proteins: Struc., Funct., Genet., 31, 74-96.
5. Avbelj F. and Baldwin R.L. (2003) Role of backbone salvation and electrostatics in generating preferred peptide backbone conformations: Distribution of phi. Proc. Natl. Acad. Sci. USA., 100, 5742-5747.
6. Avbelj F. and Baldwin R.L. (2006) Intrinsic backbone preferences are fully present in blocked amino acids. Proc. Natl. Acad. Sci. USA., 103, 1272-1277.
7. Sitkoff D. et al. (1994) Accurate calculations of hydration free energies using macroscopic solvent models, J. Phys. Chem., 98, 1978-198.

BAKER - 533 models for 99 3D/8 TR targets

Template-based Structure Prediction in CASP7 by Rosetta and Rosetta@home

B. Qian, V. Sraman, S. Khare, R. Das, W. Sheffler, D. Chivian, D. Kim, L. Malmstrom, A. Wollacott, D. Baker*
University of Washington
dabaker@u.washington.edu

CASP7 presented us with dozens of targets to test the Rosetta high resolution refinement based comparative modeling protocol we have been developing over the past couple of years. This protocol involves remodeling parts of the structures using Rosetta fragment insertion protocol, followed by Monte-Carlo minimization of the fullatom energy of the models. Our goal was to accurately model the structurally variable regions of the comparative models and to improve the structural cores over the templates. Following the steps sketched below, we improved the quality of models over the templates for many targets, but there is still considerable room for improvements.

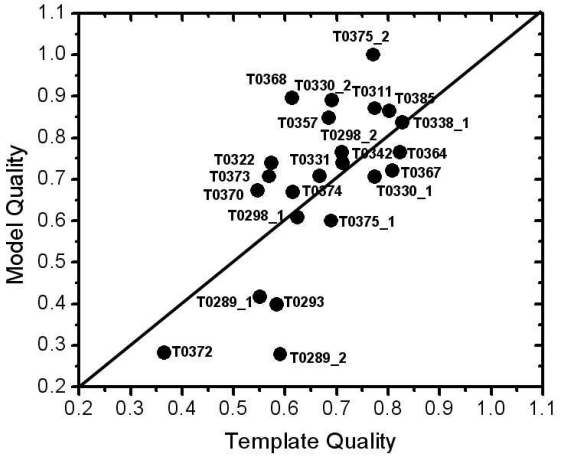
Template selection and Alignment Ensemble: The initial set of templates and target-template alignments are obtained from the 3D Jury server[1] and subjected to fullatom refinement using Rosetta fullatom energy function (see below). The templates from which the very lowest energy models were derived from are used as the candidate templates. Alignment ensemble between the candidate templates and the target sequence are parametrically generated using the K*Sync alignment method[2]. The alignment ensemble is turned into a decoy ensemble by placing the sequence of the query onto the backbone of the parent based on the alignment. Each model is then subjected to loop relax followed by fullatom refinement (see below), constrained by a set of CA-CA distance constraints generated from the decoy ensemble.

Loop Relax: To explore the conformation space around the starting comparative models, we select parts of the model that are variable from its clustered neighbors and remodeling these parts in the context of their surroundings in the starting model. The remodeling process is performed using a new loop modeling protocol, which grows loops from both ends of a loop using Rosetta fragment insertion protocol, and closes in the middle of the loops using the analytic Cyclic Coordinate Decent method[3]. Each of the structural mutants is then subjected to a number of fullatom refinements.

Fullatom Refinement: Models are refined using the Monte-Carlo minimization plus sidechain remodeling protocol described previously[4]. In each step of this protocol, a random perturbation to the protein backbone torsion angles is followed by optimization of sidechain rotamer conformations and the torsion angles flanking the site(s) of the original perturbation using the Davidon-Fletcher-Powell (DFP) algorithm. Acceptance or rejection of the new conformation is based on the Rosetta fullatom energy difference between the final minimized conformation and the initial conformation prior to the random perturbation using the Metropolis criterion. Hundreds of the above steps are preformed to obtain a low energy conformation.

Evolutionary Algorithm: Starting with full-chain structural models, we introduce structural mutations into the models using the loop relax protocol, followed by multiple instances of fullatom refinement of each mutant, and select from the population based on the Rosetta fullatom energy of each individual structure. 10 iterations are performed and the very lowest energy models are selected as submissions.

Results: In many cases we improved the quality of the starting templates with the loop relax plus fullatom refinement protocol described above. Figure 1 shows the model quality versus template quality with Maxsub 2.0Å threshold for CASP7 targets in the PSI-blast and homologous fold-recognition regimes. For the best of our five submitted models, there are a large number of cases where we have successfully improved the models over templates, with two examples illustrated in Figure 2. Falures to improve over templates are due to overly aggressive refinement, multimerization and crystal contacts in the native structures, and incorrect remodeling of the terminal segments. Possible improvements of our protocol from the insights gained during CASP7 are being pursued, with emphasis on using information from sets of homologous structures.



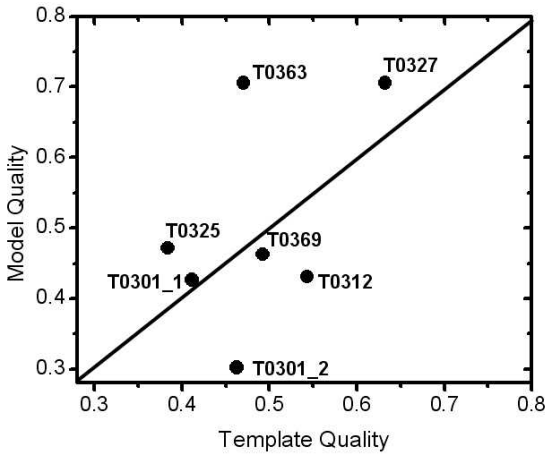


Figure 1. Quality (fraction of aligned residues) of the best models versus that of the corresponding templates with Maxsub 2.0Å threshold indicates improvement of many models over templates. The comparison is based on native structure regions that are present in the templates.

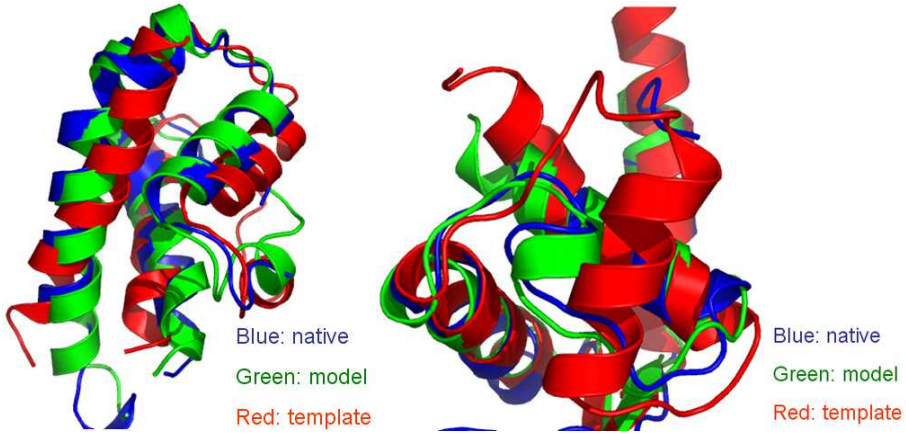


Figure 2. Superposition of the native structure(blue), the best template (red), and our best submitted model(green) for CASP7 target T0330 domain 2(left) and T0327(right).

1. Ginalski K., Elofsson A., Fischer D., & Rychlewski L. (2003) 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 19, 1015-1018.
2. Chivian D., Kim D.E., Malmstrom L., Bradley P., Robertson T., Murphy P., Strauss C.E., Bonneau R., Rohl C.A., & Baker D. (2003) Automated prediction of CASP-5 structures using the Robetta server. Proteins 53, 524-533.
3. Canutescu A.A. & Dunbrack R.L. (2003) Cyclic coordinate descent: A robotics algorithm forprotein loop closure. Protein Sci 12, 963-972.
4. Bradley P., Misura K. M., Baker D. (2005) Toward high-resolution de novo structure prediction for small proteins Science 309, 1868-1871.

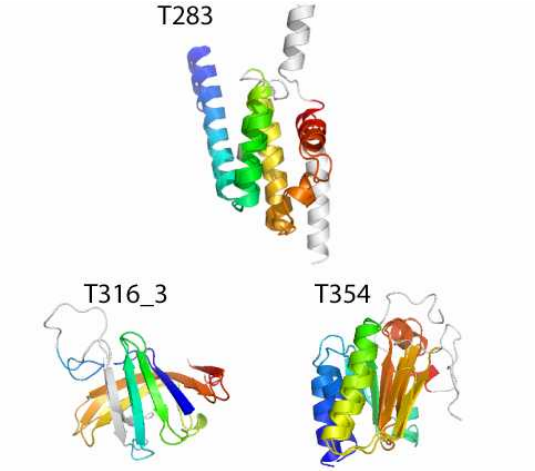
BAKER - 533 models for 99 3D/8 TR targets

Protein structure prediction by free modeling and Rosetta@home in CASP7

R. Das, R. Vernon, J. Thompson, P. Bradley, D. Bhat, M.D. Tyka, L. Malmström, and D. Baker
University of Washington
dabaker@u.washington.edu

With over two dozen new fold targets from structural genomics initiatives, CASP7 provided an unprecedented test of the Rosetta *de novo* structure prediction method. As in previous CASP experiments, we generated a large pool of 10^5-10^6 decoys by Rosetta fragment assembly[1] with a low-resolution energy function. We again attempted to ensure diversity in this decoy set by folding multiple homologs for each sequence, by forcing the exploration of different secondary structures through manually imposed torsional “bar-codes”, and by seeding simulations with long-range beta sheet pairings.[2]

Compared to CASP6, we were able to increase the computational power invested in each target sequence from 10^2 to 10^4 computer days using the distributed computing network Rosetta@home.[3] In additon to permitting an increased number of fragment insertions, the computational power was invested in the high resolution refinement of each decoy with a full-atom energy function.[4] Submitted predictions were drawn from clusters of the lowest energy decoys.



Predictions from the Rosetta full-atom ab initio method overlaid with crystal structures.

We report high resolution predictions for multiple targets, including all-alpha proteins (T283; 1.4 Å over 90 residues), all-beta proteins (domain 3 of T316; 2.8 Å over 71 residues), and alpha/beta proteins (T354; 1.8 Å over 77 residues). These successes were balanced by several cases where the Rosetta methodology did not converge well, even for small target sequences with lengths less than 100 residues. Post-mortem analysis points to numerous factors that complicated ab initio prediction of structural genomics targets: oligomerization of the proteins, highly uncertain secondary structure predictions, disordered regions, and a scarcity of sequence homologs. Solutions to these issues are being actively pursued, with strategies directly inspired by our experience in CASP7.

1. Simons K.T., Kooperberg C., Huang,E. & Baker,D. (1997) Assembly of Protein Tertiary Structures from Fragments with Similar Local Sequences using Simulated Annealing and Bayesian Scoring Functions. J. Mol. Biol. 268, 209-225.
2. Bradley P., Malmstrom L., Qian B., Schonbrun J., Chivian D., Kim D.E., Meiler J., Misura K.M., Baker D. (2005) Free modeling with Rosetta in CASP6, Proteins 61 Suppl 7, 128-34.
3. Misura K.M., Chivian D., Rohl C.A., Kim D.E., Baker D. (2006) Physically realistic homology models built with ROSETTA can be more accurate than their templates Proc Natl Acad Sci USA 103, 5361-5366.
4. Bradley P., Misura K. M., Baker D. (2005). Toward high-resolution de novo structure prediction for small proteins Science 309, 1868-1871.

BAKER-DP_HYBRID - 100 models for 100 DP targets

Hybrid domain parsing with Ginzu and RosettaDOM

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

Protein chains often contain more than one domain. In order to predict the domain organization of a protein, we have combined the Ginzu[1,2] and RosettaDOM[2] domain parsing methods into a hybrid predictor (see accompanying abstracts for Ginzu and RosettaDOM in this volume).

Ginzu attempts to determine the locations of putative domains in the query sequence and the identification of any likely homologs with experimentally characterized structures with PSI-BLAST[3] and 3D-Jury-A1[4]. This search for homologous structures is followed by parsing any remaining regions by screening Pfam[5], and then by application of a boundary preference function. The boundary preference function is derived from a PSI-BLAST[3] MSA (from the "nr" sequence database) via a heuristic that considers clusters of sequences in the PSI-BLAST MSA, the least occupied positions in the MSA, strongly predicted loop regions by PSIPRED[6], and distance from the nearest region of increased domain confidence. A fourth term boosts the likelihood of a domain boundary in regions of the MSA where the sequences frequently begin or end. Regions with structural homologs are further parsed using a consensus variant of Taylor's structure-based domain parsing method[7].

RosettaDOM generates 400 decoys structures with Rosetta's *de novo* fragment-assembly approach for the full length of the target and structurally parses each of those decoys using Taylor's structure-based domain parsing method[7]. Increased frequency of boundaries within a sliding window (smoothed in the same fashion as SnapDRAGON[8]) is used to assign domain boundaries (over a Z-score of 2.5). Although Rosetta is unlikely to produce accurate atomic-resolution models, it may accurately produce course structural features such as domains.

Both Ginzu and RosettaDOM often do not arrive at a strongly predicted boundary separately, but instead may suggest several candidate boundaries with a confidence below the threshold of each method. In such circumstances, agreement between the two methods increases the confidence of a boundary within that window. The BAKER-DP_HYBRID method takes advantage of the agreement between the sequence-based and structure based domain prediction methods by combining the boundary confidence functions from the two methods (only in regions without a strongly detected PDB homolog by Ginzu). It reports boundaries only when the combined function is above the

threshold, which may be achieved with a strong prediction by either method or when weaker predictions by each method are in agreement. Regions with PDB homologs found by Ginzu are structurally parsed with Taylor's method[7] (based on the model) in the same fashion as Ginzu.

1. Chivian D., Kim D.E., Malmstrom L., Bradley P., Robertson T., Murphy P., Strauss C.E., Bonneau R., Rohl C.A., & Baker D. (2003) Automated prediction of CASP-5 structures using the Robetta server. Proteins 53, 524-533.
2. Kim D.E., Chivian D., Malmstrom L., & Baker D. (2005) Automated prediction of domain boundaries in CASP6 targets using Ginzu and RosettaDOM. Proteins 61, 193-200.
3. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., & Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25, 3389-3402.
4. Ginalski K., Elofsson A., Fischer D., & Rychlewski L. (2003) 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 19, 1015-1018.
5. Bateman A., Birney E., Cerruti L., Durbin R., Etwiller L., Eddy S.R., Griffiths-Jones S., Howe K.L., Marshall M., & Sonnhammer E.L. (2002) The Pfam protein families database. Nucleic Acids Res 30, 276-280
6. Jones D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292, 195-202.
7. Taylor W.R. (1999) Protein structural domain identification. Protein Eng 12, 203-216.
8. George R.A., Heringa J. (2003) SnapDRAGON: a method to delineate protein structural domains from sequence data. J Mol Biol. 3

BAKER-ROSETTADOM - 99 models for 99 DP targets

The RosettaDOM Domain Parsing Protocol

D. E. Kim[1], D. Chivian[2], L. Malmström[1] and D. Baker[1]

[1] – University of Washington, Seattle, WA [2] – Lawrence Berkeley National Laboratory, Berkeley, CA

dekim@u.washington.edu

Here, we describe a protocol to identify protein domain boundaries using a sequence homology based procedure called Ginzu[1-2], and a de novo method that uses the Rosetta[3-5] structure prediction software suite for proteins lacking significant homology to experimentally determined structures.

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

1. Chivian D., Kim D.E., Malmstrom L., Bradley P., Robertson T., Murphy P., Strauss C.E., Bonneau R., Rohl C.A. & Baker D. (2003) Automated prediction of CASP-5 structures using the Robetta server. Proteins. 53 Suppl 6, 524-533.
2. Kim D.E., Chivian D., Malmstrom L., & Baker D. (2005) Automated prediction of domain boundaries in CASP6 targets using Ginzu and RosettaDOM. Proteins 61, 193-200.
3. Bradley P., Chivian D., Meiler J., Misura K.M., Rohl C.A., Schief W.R., Wedemeyer W.J., Schueler-Furman O., Murphy P., Schonbrun J., Strauss C.E. & Baker D. (2003) Rosetta predictions in CASP5: successes, failures, and prospects for complete automation. Proteins. 53 Suppl 6, 457-468.
4. Bonneau R., Strauss C.E., Rohl C.A., Chivian D., Bradley P., Malmstrom L., Robertson T. & Baker D. (2002) De novo prediction of three-dimensional structures for major protein families. J. Mol. Biol. 322, 65-78.
5. Simons K.T., Ruczinski I., Kooperberg C., Fox B., Bystroff C., & Baker D. (1999) Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins. 34, 82-95.
6. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., & Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389-3402.
7. George R.A. & Heringa J. (2002) SnapDRAGON: a method to delineate protein structural domains from sequence data. J. Mol. Biol. 316, 839-851.
8. Taylor W.R. (1999) Protein structural domain identification. Protein Eng. 12, 203-216.

BATES - 536 models for 100 3D/ 9TR targets
3D-JIGSAW - 536 models for 100 3D/ 9TR targets
3D-JIGSAW-RECOMB - 462 models for 100 3D targets
3D-JIGSAW-POPULUS - 500 models for 100 3D targets

Using genetic algorithms to recombine and refine protein models

M.N. Offman, R.A.G. Chaleil, P.W. Fitzjohn and P.A. Bates
Cancer Research UK London Research Institute
Paul.bates@cancer.org.uk

For the 7th round of CASP we used three different servers and one manual modelling procedure. Apart from our baseline protocol, 3D-JIGSAW[1], all other methods used a genetic algorithm (GA). Our overall strategy is to enhance protein modelling by considering ensembles of initial models generated from a number of different templates, alignments, scoring functions and algorithms. Since CASP6 we have developed a new GA software package called POPULUS[2]. This software adjusts protein models in internal coordinate space and is used in one of the automatic servers, for manual submissions and refinement. The GA software 'In Silico Protein Recombination'[3] is used for the third server and applies movements in Cartesian space.

Most methods for our baseline server, 3D-JIGSAW-server, have been described previously. However, for this round of CASP, five ranked models, rather than a single model, are returned where possible. The other two servers, 3D-JIGSAW-POPULUS and 3D-JIGSAW-RECOMB, base their submissions on up to 10 models created with 3D-JIGSAW and are applied five times in parallel. The final five models are ranked according to energy.

For our manual submission models, we applied an automatic pipeline using POPULUS, which has been previously evaluated for an input of approximately 200 CAFASP4 models per target[2]. For each CASP7 target, all server models were downloaded from the prediction center webpage and used as the input population. Two major changes were made to POPULUS midway through the CASP7 experiment:

First, after initialisation, for each of the models four different distance histograms are calculated[4]. These histograms represent the distances between C-alpha atoms for a sliding window of 8, 15, 22 and 29 residues. The average and standard deviation is calculated in each position. A different standard deviation cutoff has been assigned using the CASP6 submissions and the simplex optimisation algorithm for each of the four sliding windows. If sufficient "stable" distances are assigned, these averages are used as distance constraints. Comparing the four histograms of a model to the average values a penalty score can be calculated, which is used in our overall energy-scoring scheme. If there is an insufficient amount of constraints assigned the target is considered to be of the category FR/A or NF. In the latter case input models are clustered using the nearest neighbour method and only the largest two clusters are used for further progression.

Second, the movement range for mutations has been changed, to allow finer movements and to not restrict the search to the middle-points of 30*30° Φ/Ψ bins in the Ramachandran plot. Since random points within the appropriate bins are now allowed, every possible conformation within the more populated areas of the Ramachandran Plot can now be sampled.

The old and the new POPULUS protocols are each applied five times in parallel, running for at least 10 and for a maximum of 20 rounds, creating 500 models each round with a survivor rate of 10%. All ten top models are clustered. Finally a combination of energy scores, size of cluster and protein health such as buried hydrophobics, Ramanchandran Plot agreement, holes in the protein structures and g-factors are used to rank the first four models for each submission – for the protein health checks the programs QUANTA[5] and PROCHECK[6] were used. Our fifth model was submitted using only the raw energy score from the program POPULUS – with no intervention other than downloading the initial models.

1. Bates P.A. & Sternberg M.J.E. (1999) Model building by comparison at CASP3: using expert knowledge and computer automation. Proteins:Struct. Funct. Genet. 37, 47-54.
2. Offman M.N., Fitzjohn P.W. & Bates P.A. (2006) Developing a move-set for protein model refinement. Bioinformatics. 22, 1838-1845.
3. Contreras-Moreira B., Fitzjohn P.W., Offman M.N., Smith G.R. & Bates P.A. (2003) Novel use of a genetic algorithm for protein structure prediction: searching template and sequence alignment space. Proteins: Struct. Funct. Genet. 53, 424-429.
4. Carugo O. & Pongor S. (2002) Protein fold similarity estimated by a probabilistic approach based on C(alpha)-C(alpha) distance comparison. J Mol. Biol. 315, 887-898.
5. QUANTA Program (2006). Accelrys Software Inc.
6. Laskowski R.A., MacArthur M.W., Moss D.S. & Thornton J.M. (1993) PROCHECK: a program to check the stereochemical quality of protein structures. J. Appl. Cryst., 26, 283-291.

BayesHH - 100 models for 100 3D targets

Homology-based structure prediction by HMM-HMM comparison and stochastic alignment sampling

Michael Lupas[1], Johannes Söding[1]
[1]-*Max-Planck-Institute for Developmental Biology*
michael.habeck@tuebingen.mpg.de

BayesHH is one of four related servers participating in CASP7 (HHpred1 to 3, BayesHH). We originally intended to implement a fully Bayesian homology modelling step. For lack of time, we tested our alignment sampling method using MODELLER as homology modeling engine. BayesHH uses HMM-HMM comparison with integrated secondary structure comparison, correlation scoring, a novel local HMM-HMM maximum a-posteriori probability (MAP) alignment scheme, multiple template selection, intermediate profile searching, and stochastic sampling of the target-template alignment.

The tertiary structure prediction proceeds in five steps (all but step 5 are the same for HHpred3):

1. Build a multiple alignment from the target sequence with PSI-BLAST (1) (up to 8 rounds with E-value threshold 1E-3). PSIPRED (2) is used for secondary structure prediction.

2. The alignment is converted to an HMM and compared with a database of HMMs derived from representative sequences in the PDB, using the HHsearch software (3) in local Viterbi alignment mode.

3. If the top hit has a probability of less than 90% to be homologous, our intermediate profile search method HHsenser (4) is used to augment the initial target alignment.

4. The top 20 matches are clustered by UPGMA into a forest of separate trees, based on the structure comparison scores of TM-align (Zhang & Skolnick). The clustering stops when the highest average pairwise TM-score drops below 0.7. For each tree, a multiple structural alignment is calculated with MUSTANG (AS. Konagurthu et al.). The corresponding PSI-BLAST alignments are merged into a super-alignment in a master-slave fashion and an HMM is generated. The target HMM is compared with these HMMs and the best match defines a set of templates.

5. The top-scoring alignment with these templates is stochastically sampled up to 15 times. The resulting multliple sequence alignments are merged into a single target-template aligment, containing multiple instances of each template.

6. MODELLER (A. Sali et al.) is used to generate a homology model from this meta-alignment.

1. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389-3402.
2. Jones D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol. 292:195-202.
3. Söding J. (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics. 21:951-960.
4. Söding J., Remmert M., Biegert A., Lupas A.N. HHsenser: exhaustive transitive profile search using HMM-HMM comparison. Nucleic Acids Res. 2006 34:W374-8.
5. Zhang Y., Skolnick J. (2004) Scoring function for automated assessment of protein structure template quality. Proteins. 57:702-710.
6. Konagurthu A.S., Whisstock J.C., Stuckey P.J., Lesk A.M. (2006) MUSTANG: a multiple structural alignment algorithm. Proteins. 64:559-574.
7. Sali A., Blundell T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol. 1993 234:779-815.
8. http://www.ebi.ac.uk/interpro/
9. http://www.ebi.ac.uk/GOA/
10. http://scop.mrc-lmb.cam.ac.uk/scop/
11. http://www.ebi.ac.uk/interpro/

BETApro - 100 models for 100 RR targets
SVMcon - 100 models for 100 RR targets
(server, contact)

Contact Map Prediction Using BETApro and SVMcon

Jianlin Cheng and Pierre Baldi
Institute for Genomics and Bioinformatics, School of Information and Computer Science
University of California Irvine, CA 92697

In CASP7, two servers from our group BETApro [1] and SVMcon participated in contact map prediction. BETApro combines regular residue-residue contacts [2,3] with specific beta-residue contacts [1]. It improves contact map prediction for proteins containing beta-sheets. SVMcon predicts contacts (sequence separations 6) using support vector machines, integrating profiles, secondary structure, solvent accessibility, and the useful features described in [4].

1. Cheng J., and Baldi P. (2005) Three-Stage Prediction of Protein Beta-Sheets by Neural Networks, Alignments, and Graph Algorithms. Proceedings of the 2005 Conference on Intelligent Systems for Molecular Biology (ISMB 2005). Bioinformatics, vol. 21(Suppl 1), pp. i75-84.
2. Pollastri G., and Baldi P. (2002) Prediction of Contact Maps by Recurrent Neural Network Architectures and Hidden Context Propagation from All Four Cardinal Corners. Bioinformatics, vol. 18 (Suppl 1), S62-S70.
3. Cheng J., Randall A., Sweredoski M., and Baldi P., (2005) SCRATCH: a Protein Structure and Structural Feature Prediction Server. Nucleic Acids Research, vol. 33 (web server issue), w72-76.
4. Punta M. and Rost B. (2005) PROFcon: novel prediction of long-range contacts, Bioinformatics, vol. 21, pp. 2960-2968.

Bilab - 619 models for 100 3D/100 QA/ 8 TR targets
Bilab-ENABLE - 434 models for 99 3D targets

Automated tertiary structure prediction of proteins using fold recognition, model quality assessment, and fragment assembly

S. Nakamura[1], M. Kakuta[1], M. Morita[1], K. Sumikoshi[1] and K. Shimizu[1]

[1] - *Department of Biotechnology, The University of Tokyo*
shugo@bi.a.u-tokyo.ac.jp

We developed an automated protein structure prediction server named "ENABLE" and have participated in tertiary structure prediction and model quality assessment categories in CASP7. The server is based on the combination of fold recognition tools and fragment assembly method. The following is the overview of the prediction procedure of our server. 1) Search templates for the target sequence using PDB-BLAST and FUGUE[1]. 2) Execute secondary structure prediction using PSIPRED[2], disorder prediction using "disABLE", and search fragments as candidate sub-structures for each position of the target. 3) If one or more templates are found in step 1, build tertiary structure models according to the templates and alignments using MODELLER[3] and SCWRL[4]. 4) Assess qualities of generated models in step 3 using Verify3D[5,6] and determine where to improve and whether de novo prediction (start from the extended structure) is needed for the target or not. 5) Generate models from extended structure if needed (de novo prediction) or improve parts of model structures generated in step 3. 6) Pick up five models as prediction results using clustering and assessment of qualities of the models.

For predictions of disordered regions in step 2, we used disorder prediction tool named "disABLE" developed in our laboratory. This tool is based on Support Vector Machine (SVM) with position specific score matrices (PSSM) generated by PSI-BLAST[7] as input. Predictions were performed with each three different window sizes (9, 15, and 33). The weighted average of the decision values of these predictions was calculated, and disordered regions and their reliabilities were determined by these values.

For candidate fragments preparation in step 2, the length of the fragments was three to eleven, and picked up from data set generated using PISCES server [8], whose resolution cutoff was 2.5 angstrom and percentage identity cutoff was 90%, according to the similarity score including sequence identity and matching of the secondary structures. The number of candidate fragments for a target position was determined by predicted secondary structures and predicted order/disorder states: up to twenty fragments with three amino acids for disordered regions, up to twenty fragments with five amino acids for loop regions, up to twenty fragments with seven amino acids for extended regions, and twenty fragments with nine amino acids and five fragments with eleven amino acids for all regions. Redundancy of the fragments at each position was reduced according to the sequence similarity between fragments.

For model generation in step 5, we used IDDD/ABLE system based on fragment assembly method developed in our laboratory[9]. Target function including the degree of hydrophobicity of each amino acid based on predicted contact numbers [10], contacts between residues based on PSSM, average distance between hydrophobic residues, hydrogen bonds between mainchains, packing of strands, and exclusive volume to avoid overlap of residues were minimized by simulated annealing with 40000 steps.

For model quality assessment in step 6, we developed a model quality predictor based on support vector regression (SVR). Scores for a number of tools including Verify3D, ProSa[11], ProQ[12], and ABLE potential were used as inputs for SVR. Five cluster centers were picked up and submitted according to quality assessment scores.

In the case of human prediction (Bilab), initial model selection and determination of regions to be modeled in step 5 were checked and corrected by human predictor and additional models (up to 20000 structures per target) were generated in step 5 if needed.

1. Shi J., Blundell T.L. & Mizuguchi K. (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J. Mol. Biol. 310, 243-257.
2. Jones D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292, 195-202.
3. Sali A. & Blundell T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234, 779-815.

4. Canutescu A.A., Shelenkov A.A. & Dunbrack Jr.,R.L. (2003) A graph theory algorithm for protein side-chain prediction. Protein Sci. 12, 2001-2014.
5. Bowie J.U., Luthy R. & Eisenberg D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. Science 253, 164-170.
6. Luthy R., Bowie J.U. & Eisenberg D. (1992) Assessment of protein models with three-dimensional profiles. Nature 356, 83-85.
7. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W. & Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389-3402.
8. Wang G. & Dunbrack Jr.,R.L. (2003) PISCES: a protein sequence culling server. Bioinformatics 19, 1589-1591.
9. Ishida T., Nishimura T., Nozaki M., Inoue T., Terada T., Nakamura S. & Shimizu K. (2003) Development of an ab initio protein structure prediction system ABLE. Proc. 14th Int'l Conf. Genome Inform. (GIW 2003) 14, 228-237.
10. Ishida T., Nakamura S. & Shimizu K. (2006) Potential for assessing quality of protein structure based on contact number prediction. Proteins 64, 940-947.
11. Sippl M.J. (1993) Recognition of Errors in Three-Dimensional Structures of Proteins. Proteins 17, 355-362.
12. Wallner B. & Elofsson A. (2003) Can correct protein models be identified? Protein Sci. 12, 1073-1086.

BIME@NTU - 196 models for 98 DR/ 98 RR targets

DisorderPSC: Protein Disorder Prediction by Condensed PSSM, Secondary Structure, and Conservation Information

C.T. Su[1] and C.Y. Chen[2]

[1] - Department of Computer Science and Information Engineering, National Taiwan University, Taipei, 106, Taiwan, R.O.C, [2] - Department of Bio-industrial Mechatronics Engineering, National Taiwan University, Taipei, 106, Taiwan, R.O.C

sbb@mars.csie.ntu.edu.tw; cychen@mars.csie.ntu.edu.tw

Many studies have demonstrated that the disordered regions can be detected by examining the amino acid sequences. Disordered regions are distinguished from ordered regions by its low sequence complexity, amino acid compositional bias, high evolutionary tendencies, or high flexibility. In this study, a condensed position specific scoring matrix (PSSM) with respect to physicochemical properties, secondary structure, and conservation information are considered for protein disorder prediction.

In our recent work DisPSSMP[1], we investigated the predicting power of a condensed position specific scoring matrix with respect to physicochemical properties (PSSMP) on the prediction accuracy, where the PSSMP is derived by merging several amino acid columns of a PSSM belonging to a certain property into a single column. Additionally, DisPSSMP decomposes each conventional physicochemical property of amino acids into two disjoint groups which have a propensity for order and disorder respectively.

In this work, we employ a new representation for the refined SSE information and integrate it with the PSSMP features. The new representation transforms the predicted SSE information into a distance-based feature. We employ Jnet as the secondary structure predictor, which is a neural network secondary structure predictor based on multiple sequence alignment profiles[2]. Before extracting the features from the results of Jnet, a predicted SSE with less than five successive secondary structure residues are removed. We expect the remaining secondary structure segments to provide more reliable information than the original predictions. The proposed representation SSE-DIS takes the distance of a residue to its nearest secondary structure element. This feature aims to emphasize the locations which are far from the regions consisted of regular secondary structures. With the merged features of PSSMP and SSE-DIS, we invoke the QuickRBF package to construct Radial Basis Function Networks (RBFN) for classification[3]. In addition, we in particular tackle the problem of handling skewed datasets, which stands for the problems with unbalanced numbers of positive (disorder) and negative (order) samples. In order to not over-predict residues as ordered, we adopt an alternative function in determining the outputs based on the function values generated by the RBF network.

Since our training data contains more than 60% of disordered residues in terminal regions of the proteins, which causes the window-based classifiers to over-predict the terminal residues as disorder, the conservation information is considered to reduce false positives in the terminal regions. The proposed idea is based on the observation that a pair of residues are usually clustered in space and are expected to be ordered if they are simultaneously conserved. MAGIIC-PRO is an efficient pattern mining package for extracting the simultaneously conserved residues in a protein[4]. It considers large irregular gaps when growing patterns, in order to find the important residues that are simultaneously conserved but are largely apart on the sequences. In addition, MAGIIC-PRO restricts the intra-block gaps to fixed lengths, because it has been observed in previous studies that insertions and deletions are seldom present within highly conserved regions. The conservation information derived by MAGIIC-PRO is more precise than that generated by multiple sequence alignment followed by constructing the evolutionary tree.

The new predictor DisorderPSC is expected to outperform DisPSSMP after incorporating the refined information of predicted secondary structure and the concurrent conservation information with the original PSSMP features.

1. Su C.T., Chen C.Y. & Ou Y.Y. (2006) Protein disorder prediction by condensed PSSM considering propensity for order or disorder. BMC Bioinformatics 7:319.
2. Cuff J.A. & Barton G.J. (2000) Application of enhanced multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins 40: 502-511.
3. QuickRBF http://muse.csie.ntu.edu.tw/~yien/quickrbf/index.php
4. Hsu C.M., Chen C.Y. & Liu B.J. (2006) MAGIIC-PRO: Detecting functional signatures by efficient discovery of long patterns in protein sequences. Nucleic Acids Res., 34, W356-W361.

BIME@NTU - 196 models for 98 DR/ 98 RR targets

Prediction of Remote Residue Contacts by Concurrent Sequence Conservation

C.Y. Chen[1], C.T. Su[2] and C.M. Hsu[3]

[1] - Department of Bio-industrial Mechatronics Engineering, National Taiwan University, Taipei, 106, Taiwan, R.O.C, [2] - Department of Computer Science and Information Engineering, National Taiwan University, Taipei, 106, Taiwan, R.O.C, [3] - Department of Computer Science Engineering, Yuan Ze University, Chung-Li, 320, Taiwan, R.O.C.

cychen@mars.csie.ntu.edu.tw

In contact map prediction, it is considerably hard to predict remote residue contacts. While previous studies have shown that some protein residue contacts can be discovered by the occurrences of correlated mutations[1], this work expects to detect remote residue contacts by concurrent sequence conservation.

The proposed methodology is based on the secondary structure information and concurrent sequence conservation derived from sequential pattern mining. The secondary structure segments are predicted by Jnet[2], and the concurrent sequence conservation is detected by MAGIIC-PRO[4]. We call a pattern generated by MAGIIC-PRO a cluster-like pattern. The residues inside a pattern are always clustered as several sequential blocks. In between the blocks are large irregular gaps. Here comes an example: "I-x-H-N-x(52,68)-E-x(2)-L-x-K-L". In this notation, a conserved residue is recorded by its amino acid symbol, 'x' denotes an arbitrary amino acid, x(*i*) stands for a gap of *i* arbitrary residues, and x(*i*, *j*), $i < j$, represents a wildcard region of at least *i* and at most *j* arbitrary residues. This pattern contains two conserved blocks "I-x-H-N" and "E-x(2)-L-x-K". The gaps within the blocks are called intra-block gaps, and the gaps in between two sequential blocks are called inter-block gaps. Concerning the efficiency of mining process, MAGIIC-PRO specifies several constraints for these pattern components:

The maximum length of an intra-block gap: the length of intra-gap is rigid and cannot exceed the specified value.

The minimum number of residues in a block: a sequential block must contain at least a certain number of residues to eliminate noises.

The flexibility of an inter-block gap: a sequence can match a pattern as long as the inter-block gap does not violate the flexibility with respect to the query protein.

The minimum number of blocks in a pattern: a binding site is usually consisted of more than one sequential block. This constraint is set as 2 by default.

The minimum support of a pattern: the minimum percentage of sequences in the training data that match the derived pattern.

The complete procedures for discovering concurrent sequence conservation for a query protein are as follows:

Obtaining homologues of a query protein: This is achieved by running PSI-BLAST[4] against Swiss-Prot database with the BLOSUM62 substitution matrix.

Invoking MAGIIC-PRO for pattern mining: The minimum support setting is initially set as 100% and decreased repeatedly until at least one pattern with two blocks is discovered. A sequential block must contain as least three conserved residues, and the maximum length of an intra-block gap is set as 3. The flexibility of an inter-block gap is set as default.

Emerging information from different patterns: The derived patterns with exactly two blocks are collected together to calculate the conservation level of each residue. The conservation score $R(x)$ is defined by the following equation:

$$R(x) = \frac{\text{conservation level of } x}{\text{maximum conservation level among all the residues}},$$

where the conservation level of each residue is determined by the percentage of total number of supporting proteins merged from different patterns.

After the mining process completes, the contact propensity for a pair of residues *i* and *j* is defined by:

$$RR(i, j) = \sum_{i \in P, j \in P} R(i) \times R(j)$$

, where *P* is a pattern with exactly two blocks and '$i \in P$' means that residue *i* is falling in the region of one block of *P*. This information is then used to predict remote residue contacts and to differentiate paired and non-paired β-strands.

1. Fariselli P., Olmea O., Valencia A. & Casadio R. (2001) Prediction of contact maps with neural networks and correlated mutations. Protein Eng., 14(11), 835-43.
2. Cuff J.A. & Barton G.J. (2000) Application of enhanced multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins, 40, 502-511.
3. Hsu C.M., Chen C.Y. & Liu B.J. (2006) MAGIIC-PRO: Detecting functional signatures by efficient discovery of long patterns in protein sequences. Nucleic Acids Res., 34, W356-W361.
4. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W. & Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res., 25, 3389-3402.

BioDec - 66 models for 65 3D targets

All-atom Models Starting from Entropy-filtered Alignments

A. Zauli[1], I. Rossi[1]

1 - BioDec srl, Bologna, Italy

ivan@biodec.com

Here at CASP7 we blind-test the performance of a simple protocol to build all-atom models starting from for the Entropy-filtered Profile-Profile alignments [1]. This abstract summarizes the protocol used to generate the submissions for the CASP7 experiment. The following procedure is almost completely automated.

Assuming that A and B are two strings of symbols, P_A and P_B are the rectangular matrices representing the position-specific frequency of the alphabet symbols composing the strings (superscript T indicates a matrix transpose operation), S is a (symmetric) substitution matrix, it can be derived that the matrix D, defined as:

$$D = P^T_A \, S \, P_B$$

represents the "dot" matrix for the profile comparison of the two strings. This can be efficiently computed by means of standard linear algebra routines.

For each target/template comparison, we compute the dot matrix D using the composition profiles generated by multiple alignment of the sequences reported from a five-iteration PSI-BLAST [2] search on the Uniref90 database, using an inclusion threshold of $E=10^{-3}$. The scoring matrix S used S is the BLOSUM62 [3] substitution matrix. Our template set comprises the structures included in the Astral SCOP [4] database, release 1.69, whose sequence homology is less than 95%. The dot matrix D is then searched for the top scoring alignment using the global alignment with no end-gap penalties algorithm . Next, the alignments generated are subject to Shannon-entropy filtering, as described in ref. [1], using a Shannon entropy threshold of 0.5, and the remaining ones are ranked according to their Z-score. An alignment is taken into account only when its Z-score is larger than 6.

A simple cut-and-paste model is generated from the selected alignment. Side chains, non-conserved prolines and missing protein segments are then reconstructed using tools from Ram Samudrala's RAMP package (version 0.61beta) such as *scgen_mutate*, *mcgen_exaustive_loop*, and *mcgen_semfold*_loop together with the RAPDF [6] scoring function. Hydrogens are then added to the resulting model, which is then subject to energy minimization using the local BFGS algorithm as implemented in the TINKER [7] package, using the OPLS/AA [8] force field together with the GB/SA [9] implicit solvation model. The final structure is what we call the "Stage I" model.

Side-chain orientation on the "Stage I" model is then re-optimized using *scgen_double* [10] and the structure is subject to another minimization run using the same procedure described above, resulting in the "Stage II" model. The best energy-scoring structure between the "stage I" and "stage II" model is then submitted to CASP.

1. Capriotti E., Fariselli P., Rossi I., Casadio R. (2004) A Shannon Entropy-based filter detects high-quality profile-profile alignments in searches for remote homologues. Proteins 54, 351-360.
2. Altschul S.F. et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25 (17), 3389-3402
3. Henikoff S. et al. (1998) Superior performance in protein homology detection with the BLOCKS database server. Nucleic Acids Res. 26, 309-312.
4. Chandonia J. M., Hon G., Walker N.S., Lo Conte L., Koehl P., Levitt M., Brenner S.E. (2004) The ASTRAL compendium in 2004. Nucleic Acids Res. 32, D189-D192
5. http://software.compbio.washington.edu/ramp/ramp.html
6. Samudrala R., Moult J. (1998) An all-atom distance-dependent conditional

probability discriminatory function for protein structure prediction. Journal of Molecular Biology 275, 893-914,.
7. http://dasher.wustl.edu/tinker
8. Jorgensen W.L., Maxwell D.S., and Tirado-Rives J. (1996) Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. J. Am. Chem. Soc. 118, 11225-11236
9. Qiu D., Shenkin P.S., Hollinger F.P. and Still W.C. (1997) The GB/SA Continuum Model for Solvation. A Fast Analytical Method for the Calculation of Approximate Born Radii. J. Phys. Chem. A, 101, 3005-3014
10. Samudrala R., Moult J. (1998) Determinants of side chain conformational preferences in protein structures. Protein Eng. 11, 991-7.

Brooks_caspr - 108 models for 23 3D/ 7 TR targets

High-Resolution Structure Refinement Using Implicit Solvent and Replica Exchange

Jianhan Chen and Charles L. Brooks III
Department of Molecular Biology, The Scripps Research Institute
10550 North Torrey Pines Road, La Jolla, CA 92037
jianhanc@scripps.edu, brooks@scripps.edu

We have primarily focused on high-resolution refinement of server predictions in CASP7. As a blind test, it was not obvious which specific server model was the most appropriate (e.g., most native-like) for refinement. For this, we used the average potential energy during short restrained MD simulations (up to 10 ps) in a GBSW implicit solvent to roughly rank all server models for small to medium sized targets. The top 12 models with the lowest energies were compared by computing mutual CA RMSD and GDT_TS scores. The diversity, measured by average RMSD or GDT_TS values, was used as an indicator on how native-like these top models were. If the top models were believed to be sufficiently native-like (such as when average mutual RMSD is less than 3-4 Å or average mutual GDT_TS score is greater than 60-70), the model with the lowest average energy was refined using a REX/GB[1-3] protocol, described briefly below. The same protocol was also applied to refine the official CASP7 refinement targets. For many targets, the lowest energy models from short MD simulations were very diverse and no model could be reliably identified as being most native-like. We chose to carry out unrestrained REX-MD using all the top 12 models simultaneously for a few targets (using a smaller temperature range of 270-400K). Such refinement is expected to be less effective and relies almost solely on the force field to refold/refine the initial models.

The REX/GB protocol is based on all-atom replica exchange MD (REX-MD) in a generalized Born (GB) implicit solvent in CHARMM. The GB implicit solvent provides an efficient and realistic description of solvation and the REX sampling is necessary for sampling the rugged energy landscape. Furthermore, for efficacy, the sampling is focused in the vicinity of the initial models by imposing structural restraints. Without intimate knowledge of the reliable structural features, it is assumed that the initial models as selected are already native-like, such that the long secondary structure elements and the tertiary fold are largely correct. Secondary structures are enforced by dihedral restraints and the tertiary fold by residue contact derived distance restraints. All restraints are weakly imposed restraint potentials to allow a balance between stability and flexibility. We used 20 temperature replicas spanning 270-550K and the total simulation length ranged from 2.5ns to 4ns. The last 500 structures sampled at the lowest temperature (270K) during the REX simulations were clustered and the centroids were submitted as the refined models. The rank of the refined models was solely determined by the size of corresponding clusters.

It is recognized that limitations remain in the implicit solvent force field, particularly in the treatment of nonpolar solvation, which are often manifested as difficulty in modeling loosely packed structures. Additional limitation occurs in sampling capability. REX sampling improves substantially compared to simulated annealing or constant temperature simulations. Nonetheless, it is difficult to sample substantial conformation changes, which are required in some cases. Further practical complications come from oligomerization, cofactor binding and crystal packing. Such factors often have substantial impacts on the structures and lack of such knowledge can significantly hinder one's ability to refine the structures using all-atom physics-based force fields.

1. Chen J., Im W. and Brooks C.L., III. (2004) Refinement of NMR structures using implicit solvent and advanced sampling techniques. J. Am. Chem. Soc. 126, 16038-16047.
2. Chen J. and Brooks C.L. III (2006) Can molecular dynamics simulations provide high-resolution refinement of protein structure? Proteins (submitted).
3. Chen J., Im W. and Brooks C.L., III (2006) Balancing solvation and intramolecular interactions: Toward a consistent generalized Born force field. J. Am. Chem. Soc. 128, 3728-36.

CADCMLAB - 476 models for 96 3D targets

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

Carlos A. Del Carpio[1], Ismael Mohamed[1], Eiichiro Ichiishi[2], Hideyuki Tsuboi[1], Michihisa Koyama[1], Akira Endou[1], Hiromitsu Takaba[1], Momoji Kubo[1], Akira Miyamoto[1]

Graduate School of Eng. Dept. of Applied Chemistry. Tohoku University. 6-6-07 Aoba, Aramaki, Aoba-ku, Sendai 980-8579. JAPAN

We introduce a combined methodology for protein folding pattern recognition. It consists in applying a methodology for distant relative search based on our original spectral analysis methodology combined with more orthodox sequence alignments techniques. The concept behind the spectral analysis method is a periodicity analysis of the physicochemical properties of the residues constituting proteins primary structures. The analysis is performed using a front-end processing technique in automatic speech recognition[1,2] by means of which the cepstrum (measure of the periodic wiggliness of a frequency response) is computed that leads to a spectral envelope that depicts the subtle periodicity in physicochemical characteristics of the sequence. A diversity of proteins are extracted when this methodology is applied to the search of similar protein folding patterns to a particular target. Extracted structures rank from scant similarity in terms of amino acid composition to high similarity ones. Then a more specific sequence alignment (like FASTA or BLAST) can be applied to the reduced set of structures obtained by our spectral oriented methodology. This combined method has shown a high degree of effectiveness to select optimal templates for a determined target, both in terms of processing times as well as quality of template. The threading algorithm is then pursued by an energy minimization process for the newly built structure.

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

CaspIta-FOX - 499 models for 100 3D targets

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

P. Fontana[1], F. Sirocco[2], S.C.E. Tosatto[2], R. Velasco[1] and S. Toppo[3]

[1] - Istituto Agrario di San Michele all'Adige
[2] - Dip. di Biologia & CRIBI Biotech Centre, Universita' di Padova
[3] - Dip. di Chimica Biologica, Universita' di Padova
stefano.toppo@unipd.it

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

In the first phase a secondary structure prediction of the target is performed based on the ConSSPred[1] protocol. This prediction is used to search for hits against a database of known secondary structures extracted from PDB (using DSSP) by means of a global alignment search based on SSEA[2] (Secondary Structure Element Alignment) available at the following website http://protein.cribi.unipd.it/ssea. At the end of the first phase a list of hits that share a similar secondary structure topology with the target sequence is extracted.

The second phase is based on a modified protocol for scanning the sequence database called SENSER[3]. A procedure based on four iterations against NR60 database and the last one vs PDBAA is used to identify a template structure for the target sequence. NR60 is produced by applying the CD-HIT[4] algorithm to cluster the NR database at 60% sequence identity. Once putative templates are found, they are back validated. The back-validation step consists in using PSI-BLAST[5] to find the target starting from a different query sequence. I.e. due to the asymmetric nature of PSI-BLAST, if sequence A finds sequence B it is not always the case that B also finds A. Sequences that back-validate are more likely to be correct hits even at low sequence similarity. If no significant hit is found, or the hit does not back-validate, a new PSI-BLAST search, using the above "4+1" protocol on NR60 and PDBAA, is started for the highest ranking sequences (i.e. lowest e-value) belonging to the sequence space or profile of the target sequence. Once a sequence from PDBAA back-validates and its secondary structure is compatible with the one of the target sequence as found in the first phase, the protocol builds a target to template alignment and stops.

In order to produce an accurate alignment, a profile-profile alignment approach has been used. The method is based on a program developed for the Arby

server[6] which uses information from secondary structure predictions and sequence profiles. Alignments are automatically generated by systematically testing 625 different parameter combinations involving the weigths given to sequence profile and secondary structure of both target and template. Five values of each parameter are tested and chosen from a reasonable range[7]. Each target-template alignment is used to build a raw model whose quality is evaluated on the basis of its estimated quality[8]. The best scoring target-template alignment is chosen to build and refine the final model.
The final model is generated using the package HOMER (http://protein.cribi.unipd.it/Homer). This involves the following steps. First a raw model of the conserved parts is constructed from the template. The conserved backbone 3D coordinates are copied and missing side chains placed with SCWRL[9]. Insertions and deletions are reconstructed using an enhanced version of the fast divide & conquer loop modeling method[10]. An experimental version of the FOX server is available at the following website address http://protein.cribi.unipd.it/fox.

1. Albrecht M., Tosatto S.C.E., Lengauer T. and Valle G. (2003) Simple consensus procedures are effective and sufficient in secondary structure prediction. Protein Engineering, 16, 459-462.
2. Fontana P., Bindewald E., Toppo S., Velasco R., Valle G. and Tosatto S.C. (2005) The SSEA server for protein secondary structure alignment. Bioinformatics 21, 393-395.
3. Koretke K.K., Russell R.B. and Lupas A.N. (2002) Fold recognition without folds. Protein Science, 11, 1575-1579.
4. Li W., Jaroszewski L. and Godzik A. (2002) Tolerating some redundancy significantly speeds up clustering of large protein databases. Bioinformatics. 18, 77-82.
5. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W. and Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Acid. Res. 25, 3389-3402.
6. Von Ohsen N., Sommer I., Zimmer R. and Lengauer T. (2004) Arby: automatic protein structure prediction using profile-profile alignment and confidence measures. Bioinformatics. 20, 2228-2235.
7. Sommer I., Toppo S., Sander O., Lengauer T. and Tosatto S.C.E. (2006) Improving the quality of protein structure models by selecting from alignment alternatives. BMC Bioinformatics. 27, 364.
8. Tosatto S.C.E. (2005) The victor/FRST function for model quality estimation. J Comput Biol. 12, 1316-1327.
9. Canutescu A.A., Shelenkov A.A. and Dunbrack R.L.Jr. (2003) A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci. 12, 2001-2014.
10. Tosatto S.C.E., Bindewald E., Hesser J., Manner R. (2002) A divide and conquer approach to fast loop modeling. Protein Eng. 15, 279-286.

CaspIta-FRST - 93 models for 93 QA targets

The Victor/FRST Function for Model Quality Estimation

S.C.E. Tosatto[1]

[1] – Dept. of Biology and CRIBI Biotech Centre, University of Padova
silvio@cribi.unipd.it

The Victor/FRST[1] (Function of Rapdf, Solvation and Torsion potentials) function is a statistical scoring function used to estimate the quality of a protein structure. It is implemented as the weighted linear combination of four different components covering the major aspects of structure quality estimation.

The first component is an implementation of the RAPDF[2] statistical pairwise potential. This potential of mean force discriminates between residue specific non-bonded interactions at the atomic level, e.g. the C_α of an Isoleucine is a different type from the C_α of a Glycine. It is used with published parameters. A simple solvation potential is derived in analogy to the one described for GenTHREADER[3]. The relative solvent accessibility is estimated as the number of other C_β atoms within a sphere of radius 10 A centered on the residue's C_β atom. The reference state for this distribution is generated from the TOP500 database[4]. This database of high resolution crystal structures is used to estimate the relative probability of encountering a number i (i = 0,…,40) of C_β atoms surrounding each of the 20 amino acids. The energy for a given structure is calculated with the standard log scale for mean force potentials. A similar scheme was also used to parameterize the torsion angle potential. All (ϕ,ψ) angle combinations, discretized in 10x10 degree bins, present in the TOP500 database[4] are used to estimate the reference state for each of the 20 amino acids. The same log scale formula is applied to derive an energy for a given structure. Finally, a crude hydrogen bond potential was derived by counting the number of backbone N – O pairs falling within a given distance cutoff[1].

Since the four components have different orders of magnitude and cannot be related directly to the same scale, weighting factors are used before summing the partial energies. These factors were optimized on the CASP-4 decoy set[5] optimizing the linear correlation between total energy and GDT_TS score[6] as target function. The final scoring function was used to submit QA predictions to CASP-7.

1. Tosatto S.C. (2005) The Victor/FRST Function for Model Quality Estimation. J Comput Biol, 12, 1316-1327.
2. Samudrala R., & Moult J. (1998) An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J. Mol. Biol., 275, 895-916.

3. Jones D.T. (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J. Mol. Biol., 287, 797-815.
4. Lovell S.C., Davis I.W., Arendall W.B.r., de Bakker P.I., Word J.M., Prisant M.G., Richardson J.S., & Richardson D.C. (2003) Structure validation by Calpha geometry: phi,psi and Cbeta deviation. Proteins, 50, 437-450.
5. http://predictioncenter.llnl.gov/download_area/CASP4/MODELS_SUBMITTED/
6. Zemla A. (2003) LGA: A method for finding 3D similarities in protein structures. Nucleic Acids Res., 31, 3370-3374.

CaspIta-FRST-SVM - 98 models for 98 QA targets

FRST-SVM: Predicting Model Quality from Statistical Potentials and Structural Features Using Kernel Machines

S.C.E. Tosatto[1], A. Vullo[2], G. Pollastri[2]
[1] – *Dept. of Biology and CRIBI Biotech Centre, University of Padova,*
[2] – *School of Computer Science and Informatics, University College Dublin*
silvio@cribi.unipd.it

FRST-SVM is an extension of the Victor/FRST[1] (Function of Rapdf, Solvation and Torsion potentials, see accompanying abstract for group CaspIta-FRST) function for protein structure quality estimation. Unlike its predecessor, it uses a support vector machine (SVM) to combine partial scores covering the major aspects of structure quality estimation. Several additional features describing the structure under scrutiny were also added compared to the previous version.

The features used to train the SVM include the four previously described statistical potentials used for FRST[1] and include pairwise, solvation, hydrogen-bonding and torsion angle terms. A normalized torsion angle propensity derived from scoring the model against the maximum attainable torsion angle score was added together with ten structure based features. The latter represent the length of the protein structure, its fraction of secondary structure (α, β, coil) content and hard sphere backbone $C_\alpha - C_\alpha$ clashes at less than 2.75 A distance. Five features are a count of $C_\alpha - C_\alpha$ chain breaks (distance > 4.5 A) at increasing distance thresholds (< 7.5, < 10, < 15, < 20, > 20 A).

SVM training was performed using the LIBSVM package[2] with a radial basis distribution function. Nearly 4,000 models from the CASP-4 decoy set[3] were used as training set for cross-validation experiments. The SVM was trained in regression mode in order to predict the TMscore[4] of each decoy structure. The final scoring function was used to submit QA predictions to CASP-7.

1. Tosatto S.C. (2005) The Victor/FRST Function for Model Quality Estimation. J Comput Biol, 12, 1316-1327.
2. Chang C.C. and Lin C.J. (2001) LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm
3. http://predictioncenter.llnl.gov/download_area/CASP4/MODELS_SUBMITTED/
4. Zhang Y. and Skolnick J. (2004) Scoring function for automated assessment of protein structure template quality. Proteins, 57, 702-710.

CaspIta-GOret - 227 models for 100 FN targets

GOretriever (Gene Ontology retriever): a fast automated protein function annotation based on semantic similarities

P. Fontana[1], A. Cestaro[1], L. Demattè[1], R. Velasco[1] and S. Toppo[2]
[1] - *Istituto Agrario di San Michele all'Adige*
[2] - *Dip. di Chimica Biologica, Universita' di Padova*
stefano.toppo@unipd.it

We present a method, GOretriever, for fast annotation of protein functions based on Gene Ontology (GO) terms[1] clustering. Presently the method is based on two distinct phases to recover a putative function for the target protein. The first step is based on a PSI-BLAST[2] search against UniProt[3] database of annotated proteins. The second phase is a clustering procedure of GO terms that belong to the found hits.

The Gene Ontology[1] (GO) is based on a structured vocabulary of protein functions where each term is described as a father-child relationship and multiple inheritances are allowed. In this framework protein functions are represented by a DAG (Directed Acyclic Graph) starting from the root, consisting of general terms, to the leafs containing different levels of detailed descriptions. Such an ordered infrastructure makes feasible to infer and measure semantic similarities of distant or different concepts simply looking at the information content they share.

In its present form, the tool is based on a five iterations PSI-BLAST search vs. the UniProt database to extract related proteins. We have used the default searching cutoffs and increased the number of hits to show up to 1000 to assess the statistical measures.

The GO terms extracted from the hits are processed in order to reconstruct all of the possible paths that lead to the root node. During the recursive process

each node is scored adding the weights of the nodes encountered during the path reconstruction. The weights of the nodes depend on the scores of the hits found by the PSI-BLAST search. As a result we obtain a trimmed GO graph consisting only of the terms found in the database search: for each term we keep track of its occurrence and of its cumulative score.

Since the most frequent nodes are the least informative ones, as we get near the root, the algorithm tries to find a good balance between the occurrence, weight of a node and its measure of information content in order to find the most probable paths. Since these nodes may still be highly spread, a clustering approach has been used. GO terms are tentatively grouped on the basis of their Information Content (IC) and their semantic distances calculated applying the Lin formula[4] that computes the amount of information shared. In this phase only the most informative GO term is retained as group representative. The final list of filtered and retained hits are then ranked efficiently using two statistical scores and an entropy based measure: "Internal Confidence" (InC), "Absolute Confidence" (AC) and Theil Index (TI)[5]. The InC and AC scoring methods have been specifically developed to assess the statistical significance of the retrieved hits and are both based on non-cumulative node weights divided by either cumulative root node weight (InC) or by the maximal theoretical weight (AC). Theil index (TI) is derived from Shannon's measure of information entropy[5] and it is applied to measure the inequality of score distribution over the trimmed GO graph. The program output is a list of ranked GO terms with the highest score and information content (IC).

1. Harris M.A., Clark J., Ireland A., Lomax J., Ashburner M., Foulger R., et al. (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 32: D258-61.
2. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W. & Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389-3402.
3. Bairoch A., Apweiler R., Wu C.H., Barker W.C., Boeckmann B., Ferro S., Gasteiger E., Huang H., Lopez R., Magrane M., Martin M.J., Natale D.A., O'Donovan C., Redaschi & N., Yeh L.S. The Universal Protein Resource (UniProt). Nucleic Acids Res. 33:D154-159.
4. Lin D. (1998) An Information-Theoretic Definition of Similarity. Proceedings of the Fifteenth International Conference on Machine Learning. 296-304
5. Henri T. (1979) The measurement of inequality by component of income. Economics Letter 2.

CBiS - 15 models for 4 3D targets

A new ab initio mathematical model for protein structure predictions

Yi Fang[1], Junmei Jing[2]

[1]-Centre for Bioinformation Sciences, The Australian National University , [2]-Centre for Bioinformation Science, The Australian National University

yi@maths.anu.edu.au

There are 4 well-known structural features of native structures of globular proteins:

- High density;
- Small surface area;
- Hydrophobic core;
- Long peptide chains fold into domains.

To form an ab initio mathematical model based on these features we make a working hypothesis:

Under complicated combined physico-chemical forces, in the physiological environment nature pushes a globular protein to form a conformation which is compactly packed, and simultaneously satisfies the above features in a cohesive way.

We put a conformation P of a peptide chain U into a tailor-made closed thermodynamic system S(P). Then the complicated physico-chemical interactions reduce to boundary conditions of the system. We translate the above features into 3 ariables: system volume V(P); system boundary area A(P); system boundary hydrophobic area W(P). The above features show that the smaller these quantities are, the higher the density and the better the hydrophobic core will form. Thus by minimizing an energy function

$$u\_n(P) = E\_n(V(P), A(P), W(P))$$

among meaningful conformations, instead of all conformations, the model imitates nature by pushing the peptide chain into a conformation that best suits the above features. The E_n is an increasing function for each of its 3 variables and n is the chain length. Due to the fourth feature, the energy function may depend on n. By the hypothesis, a minimizing conformation is the native structure.

We use an all-atom space-filling model to represent a conformation. The meaningful conformations are defined by steric conditions that reflect effects of complicated physico-chemical interactions, except for the hydrophobic interaction which is reflected by the energy function. The steric conditions

restrict minimum atomic distances in a conformation and avoid inaccuracies in approximated energy calculations by passively relying on well-known geometric restrictions of protein natve structures. For bonded atoms, the allowed distance is around the standard bond length. For a pair of non-bonded atoms, their physical-chemical properties in the molecule decide the minimum distance. For example, atoms with different charges have smaller minimum distance allowed than that for the same charged ones; sulphur atoms in different Cysteins have minimum distance allowing disulfide bond. The steric conditions play an equally important role as the energy functions, only with them the model can distinguish very similar peptide chains such as a wild chain and a one-residue mutation.

We used the molecular surface as system boundary and linear functions, or weighted averages of A(P), W(P) and the 2/3rd power of V(P) as energy functions. Mathematically these are good approximations. We found that secondary structures and hydrogen bonds, although never pursued by the model, always appear in our predicted structures. These results partially verify the working hypothesis since the appearance of secondary structures and hydrogen bonds is an inference from the hypothesis.

Our prediction program uses the gradient method. Since the gradient program was not ready until August 4th we only predicted the last several targets. From an extended conformation, we rotate all rotatable bonds in one round according to the gradient. Then we check the steric conditions for the new conformation. If not satisfied, we reduce the length of the gradient and try again. Continuing, we either achieve a conformation that has zero gradient, or one for which any tiny rotation around any rotatable bond will violate the steric conditions. In either case, we have finished a run and record the structure. Then we will start next run with a random change of the extended conformation. Using our digital machine of 730 Mhz processor, each run needs about one hour. Accumulating several runs, we select the best results as the predicted model. The most time consuuing part is checking steric conditions. Due to bugs in our rotation and checking programs, we omitted the checking to make the deadline. Since the energy function counts only the hydrophobic interaction, this omission shows that hydrophobic interaction alone produces secondary structures.

In later runs we get molting globes for various linear energy functions. With deeper study about energy functions and better programming, this model has very good potentials.

CBRC-DP_DR - 200 models for 100 DP/ 100 DR targets

Prediction of disordered coil regions in proteins by fold recognition and secondary structure prediction

T. Noguchi[1], M. Takizawa[1,2], N. Inoue[3] and K. Tomii[1]
[1] – *Computational Biology Research Center*
National Institute of Advanced Industrial Science and Technology, Japan
[2]- *Graduate School of Science & Engineering, Waseda University, Japan*
[3] –*Pharma Design, Inc., Japan*
noguchi-tamotsu@aist.go.jp

We predicted structurally disordered coils in protein sequences using a protocol based on the following three steps: 1) We identified putative coil regions using fold recognition methods or secondary structure predictions; 2) We calculated the disorder propensity of the putative loop regions identified above. 3) Finally, we checked that the above predicted disordered regions were not inter-domain regions using domain linker prediction programs. This method was succeeded in predicting the domain boundary in CASP6. We have updated our method by using better fold recognition methods for CASP6 and the secondary structure prediction methods, which are better prediction accuracy. We used FORTE1 [1], FUGUE2 [2], FFAS03 [3] and SAM-T02 [4] for fold recognition, and PSIPRED [5], NSSP [6], SSpro [7], Prof [8] and SAM-T02 for secondary structure prediction. DLP [9] and DomCut [10] are used for the domain linker prediction, which were used at CASP6.

For CASP7 targets, we have prepared the three different methods, based on single fold recognition method (FORTE1), the consensus of three fold recognition methods (FUGUE2, FFAS03 and SAM-T02) and the consensus of the above five secondary structure predictions, to identify the coil regions.

In step 1, loop regions were determined using the fold recognition method or the secondary structure prediction. For the methods based on the fold recognition, we identified the coil regions of a target sequence by a single or a consensus alignment on the template structure. When the template structures differed among the three fold recognition methods, the alignment on the template with SAM-T02 was used. For the method based on the secondary structure prediction, consensus secondary structure predictions were used to identify coils in regions. We prioritized predictions of PSIPRED, when no consensus secondary structure prediction was obtained.

In step 2, we predicted disordered loop regions in proteins using the propensity and the loop regions as defined above, and according to the following criteria. All coil regions with three or more consecutive amino acids with high propensity and with an average propensity greater than 1.2 were predicted to be structurally disordered.

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

The results by three methods were carefully analyzed with reference to the template structure and/or the predicted structure, and the final disordered regions and domains were determined.

1. Tomii K. & Akiyama Y. (2004) FORTE: a profile-profile comparison tool for protein fold recognition. Bioinformatics, 20, 594-595.
2. Shi J., Blundell T.L., Mizuguchi K. (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol, 310(1), 243-57.
3. Jaroszewski L., Rychlewski L., Li Z., Li W., Godzik A. (2005) FFAS03: a server for profile-profile sequence alignments. Nucl. Acids Res. 33, W284-W288.
4. Karplus K., Karchin R., Draper J., Casper J., Mandel-Gutfreund Y., Diekhans M., Hughey R. (2003) Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. Proteins, 53 Suppl 6:491-6.
5. McGuffin L.J., Bryson K., Jones D.T. (2000) The PSIPRED protein structure prediction server. Bioinformatics, 16,404-5.
6. Salamov A.A., Solovyev V.V. (1995) Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiply sequence alignments. J.Mol.Biol., 247, 11-15.
7. Pollastri G., Przybylski D., Rost B. & Baldi P. (2002) Improving the Prediction of Protein Secondary Structure in Three and Eight Classes Using Recurrent Neural Networks and Profiles. Proteins, 47, 228-235.
8. Ouali M., & King R.D. (2000) Cascaded multiple classifiers for secondary structure prediction. Prot. Sci.,9, 1162-1176.
9. Miyazaki S., Kuroda Y. & Yokoyama S. (2002) Characterization and prediction of linker sequences of multi-domain proteins by a neural network. Journal of Structural and Functional Genomics, 2, 37-51.
10. Suyama M. & Ohara O. (2003) DomCut: prediction of inter-domain linker regions in amino acid sequences. Bioinformatics, 19, 673-674.

CBRC-DR - 100 models for 100 DR targets

POODLE: predicting protein disorder using machine-learning approaches

K. Shimizu[1], S. Hirose[2], N. Inoue[2], S. Kanai[2] and T. Noguchi[1]

[1] - Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology, Japan

[2] - Pharma Design, Inc., Japan

poodle@cbrc.jp

We predicted protein-disordered regions using machine-learning approaches. We prepared three types of disordered prediction (POODLE-S, L and W) according to the length of the target disorder.

POODLE-S version puts emphasis on predicting short disorder regions[1]. Because the amino acid composition has different propensities in the N-term, C-term, and internal regions, the accuracy of prediction increases by dividing training data into several regions and by predicting them separately. We calculated the chi-square scores over all pairs of ten-residue windows for every five residues. Then, we separated the data using a chi-square score with a 5% significance level so that each data item had the same amino acid compositional tendency. Also, each defined region has different physico-chemical properties (hydrophobic, positive, negative, charged, polar, small, tiny, aliphatic, aromatic), which are important factors contributing to disorder. We selected specific features for each region. The method for POODLE-S has three steps. In the first step, PSSMs of target sequences are calculated via PSI-BLAST. In the next step, PSSMs are divided into sliding windows of size m (If the windows are on terminal areas←unclear, m=5. If not, m=15.). Then, each window is $m \times n$ matrix $\boldsymbol{E}_{i,j}$ {i=1…m, j=1…*20*} (j shows 20 types of amino acids). In the last step, features are extracted from windows. Each feature, $\boldsymbol{F}_{i,c}${i=1…m, c=1…f}(f shows the number of selected features for each region), is calculated as follows. $\boldsymbol{F}_{i,c} = \boldsymbol{E}_{i,j}$ (if j has characteristic c). Then, each extracted feature is classified into disorder or order using support vector machines (SVM)[2].

POODLE-L version puts emphasis on predicting long disorder regions, mainly ones longer than 40 consecutive amino acids. POODLE-L was a set of disorder region prediction models. Each prediction model consisted of a two-step prediction using SVM. In the first step, the model predicted whether the sequence of 40 consecutive amino acids in the window was disordered or not, based on ten physico-chemical descriptors. In the second step, it predicted whether each residue was disordered or not, based on the distribution of probabilities obtained in the first step. To start with, the model was designed using the ten descriptors in the first step, which was called the original model. Next, 62 models were created by changing six descriptors groups, into which the ten parameters in the step were classified based on the physico-chemical

properties of amino acid. The prediction accuracy of these models was then compared with that of the original model, and nine models with higher performance than the original model were selected. POODLE-L integrated the prediction results of the models and the original model by adopting a regional consensus as follows. Windows with 7, 25, and 39 residues were set for every prediction result created by the ten models. For the prediction results of each model, the mean value of probability in each window was then calculated. The results were sorted in a large order, and two top and bottom values were then removed. Consequently, six probabilities existed in each window. The average score of 18 mean values of probability obtained in each window was finally assumed to be the result of the disorder prediction in the center of the amino acid in the window.

POODLE-W is a binary predictor, which classifies a target protein to be mostly folded or disordered. POODLE-W was developed to avoid training data bias using a semi-supervised learning approach because few disordered proteins are available. POODLE-W uses a spectral graph transducer[3] that utilizes the information on structure-known proteins as well as the information on structure-unknown proteins.

POODLE-S, L and W are trained on sequences that are extracted from PDB and DisProt. POODLE-W also uses SwissProt for training. All information about the POODLE series is provided at http://mbs.cbrc.jp/poodle/ .

1. Shimizu K., Muraoka Y., Hirose S. & Noguchi T. (2005) Feature Selection Based on Physicochemical Properties of Redefined N-term Region and C-term Regions for Predicting Disorder. Proceedings of 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, pp. 262-267.
2. Vladimir N. & Vapnik V. (1995) The Nature of Statistical Learning Theory. Springer.
3. Joachims T. (2003) Transductive Learning via Spectral Graph Partitioning. Proceedings of International Conference on Machine Learning 43-151.

CBSU - 287 models for 100 3D targets

Protein Structure Models based on Fold-Recognition Templates and Their Remote Structural Neighbors

D. R. Ripoll

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

ripoll@tc.cornell.edu

We developed a protein structure prediction approach that was systematically applied to all the CASP7 targets. The main source of structural information to model each target was collected from a series of automatic servers, such as the BIOINFO (3D-Jury)[1], ROBETTA[2] and LOOPP[3]. The templates used in the structure generation of our models were selected as follows: (i) They corresponded to top-score, high-confidence predictions identified by the servers we used; (ii) if there was no consensus among the servers or the predictions were given a low-level of confidence, then, we used templates for which most of the secondary structure elements were arranged linearly as in the secondary structure predictions of the target sequence. (iii) For those targets for which the servers where not able to assign high confidence scores to any template and the secondary structure predictions did not match the sequential arrangement of α-helices and β-strands, alternative templates were built by permutations in the sequential order of the secondary-structure elements of low-confidence templates from the servers list.

In addition, attempts were made to improve the predicted sequence alignments provided by the servers in those cases where no obvious homology was detected. To achieve this objective, structural alignments of the template structure with proteins sharing the fold but having low sequence identity were used to help identifying the *essential* secondary-structure elements specific to the fold and the regions of high sequence and/or structure variability. Structural neighbors of the templates were identified using the Combinatorial Extension methodology of Shindyalov and Bourne[4]. The commercial program ICM-Pro (Molsoft, Inc) was also used for checking visually the pairwise assignments in the structural alignments of templates and their structural neighbors. The DS-Modeling program (Accelrys Inc) was subsequently used in attempts to optimize the initial alignment from the servers by using the structural alignment of template and neighbors mentioned above, and properties such as conservation of hydrophilic/hydrophobic residues in the experimental structures. All-atoms 3D models for the targets were generated by using the programs MODELLER[5] or ECEPPAK[6] and inspected visually for consistency. A set of rules were systematically applied to all model: (a) putative fragment deletion in the target sequence cannot eliminate a central strand from a β-sheet; (b) an insertion falling inside an α-helical region was, either shifted toward the

nearest loop region in the template fold, or forced into an α-helical conformation; (c) the insertions falling in the middle of a β-strand were shifted toward the nearest loop region.

1. Ginalski K., Elofsson A., Fischer D., Rychlewski L. (2003) 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics. 19, 1015-1018; (http://bioinfo.pl/meta/).
2. Simons K.T., Kooperberg C., Huang E., Baker D. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol. 268, 209-225.
3. Teodorescu O., Galor T., Pillardy J. and Elber R. (2004) Enriching the sequence substitution matrix by structural information. Prot., Struc, Funct. Bioinform., 54, 41-48.
4. Shindyalov I.N., Bourne P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Engineering 11, 739-747.
5. Šali A., Blundell T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234, 779-815.
6. Ripoll D.R., Liwo A., Czaplewski C. (1999) The ECEPP Package for Conformational Analysis of Polypeptides. TASK Quart. 3, 313-331.

CDAC - 4 models for 4 3D targets

Hybrid Methods for Predicting the Protein Structures

V. Sundararajan and Swapna Gunda

Scientific and Engineering Computing Group
Center for Development of Advanced Computing
Pune University Campus. Pune-411 007, India.
vsundar@cdac.in, swapnag@cdac.in

In the post genomic era, with the explosion of protein sequence data, there is a need for understanding the structure of protein in order to elucidate their function. Since experimental techniques cannot meet this challenge, theoretical methods are required. Commonly used, knowledge based and *ab initio* approaches are showing great promise in high-resolution structure prediction, albeit their own pros and cons. So an *ab initio* model is developed using Genetic Algorithms (GA), which is based on the search for better structures in a torsion angle space. PSP is posed as a minimization of energy function with torsion angles as the basic variables. The variables are coded in a binary string to represent 360 degrees variations and each angle is represented by binary code of 9 bits. Genetic operations are carried over on a population of binary strings of ϕ,ψ(back bone dihedral angles) and χ_1(side chain dihedral angle) with force field including van der Waals, electrostatic, hydrophobic and torsion angle interactions. A pseudoentropic term is added to preclude extended structure formation. The best individual is judged based on the energy being minimum. Tournament selection is used with a selection pressure of two, followed by crossover and mutation probability and this process is repeated. This constitutes the simple GA, which is being used by the current method as the base.

All simulations were performed using Fortran 90 under Unix environment. The internal coordinates required in building the molecule and the potential energy function is taken from force field AMBER94. In the interactive process of GA the most time consuming part of the calculation is the individual fitnesses. To reduce this time complexity a data parallel model is done by a master-slave approach, which distributes the calculation of fitnesses to different processors in every generation. This is developed with MPI standard and run using PARAM, a series of supercomputers developed by C-DAC implementing open frame architecture. The code is portable to any other parallel machine.

Since the number of conformations accessible to a polypeptide chain grows exponentially with chain length, the logical starting point for the development of models attempting to describe the folding of real protein is testing on very small proteins of known structures. So initially the model is tested for peptides of residue length<15 including Octaalanine, Alcohol Dehydrogenase, Citrate Synthase and Troponin-C and observed to predict well with RMSD<3Å. But the results were not satisfied if peptide length increases showing higher RMSD and also short-contacts. So a new strategy, named Divide & Evolve method has been implemented based on the hypothesis that, “divide the polypeptide into smaller fragments, predict each independently using simple GA, and evolve as a whole again by varying only connecting angles between fragments using Monte Carlo steps”. This method was tested for Villin HP-36, Crambin and Amyloid beta. They were predicted with RMSD between 5 and 9Å with some of the secondary structure elements matching the experimental ones. The same method is applied for predicting CASP targets T0348 (68 residues), T0335 (85 residues), TO358 (87 residues) and TO359 (97 residues). The structures were predicted with out any geometrical inconsistencies. From the simulations of test-set proteins it was able to conclude that: (i) Addition of variation in side chain dihedral angle χ_1 to back bone dihedral angles (ϕ and ψ) has shown improvement in the results. (ii) The structure represented by the average torsion angles of all minimum conformations obtained in simulation is observed to be closer to the experimental result than the final minimum energy structure.

1. Sundararajan V.(2003) Predicting Three Dimensional Protein Structure Using Genetic Algorithms: A review, IICAI-03, pp.668-675.
2. Sachin B Karale, Jayaraman V.K., Swapna Gunda and Sundararajan V. (2005) Ant Colony Approach To Bio-molecular Structure Optimization, International Conference on Computational & Experimental Engineering and Sciences, ICCES-05.

Chen-Tan-Kihara - 653 models for 97 3D/ 34 QA targets

Fold recognition prediction based on suboptimal DP

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

Current fold recognition methods usually build the predicted model based on the optimal structure-sequence alignment. However, by using suboptimal alignments, we can extract more information for structure prediction. From this idea, we developed our approach, named SUBWAy, for the CASP7 structure prediction category, which combines several suboptimal alignments to build the final model.

The fold recognition method is established on the suboptimal dynamic programming algorithm[1]. The scoring scheme in the algorithm is a weighted combination of a profile-profile alignment and secondary structure information. The profile-profile alignment is based on PSIC (position-specific independent counts) weighting[2], PSI-BLAST pseudocount[3] and symmetric log-odds multinomial score[4]. The scoring matrix for secondary structure correspondence is taken from a paper by Dunbrack et al[5].

The template structure pool for this method consists of 4600 non-redundant protein structures which are filtered out from the PDB-REPRDB[6].

In the template recognition phase, the whole template pool is scanned for every target. Five templates with the highest alignment scores are picked up for model construction. If templates used by some other CAFASP models have a higher alignment score, those templates were also considered.

In the model construction phase, every recognized template is aligned with the target. The top 5 suboptimal alignments are fed to Modeller[7] to build up an average 3-D structure model. We check if the predicted model contains some inter-Cα clashes. If a clash is found, a distance restriction was added to Modeller to eliminate this clash and rebuild the model. This process was iterated until no Cα clash exists in the predicted model. These refined models are ranked by their reliability score which reflects the confidence of the model for final submission.

1. Vingron M. & Argos P. (1990) Determination of reliable regions in protein sequence alignments. Protein Eng. 3, 565-569.
2. Sunyaev S.R., Eisenhaber F., Rodchenkov I.V., Eisenhaber B., Tumanyan V.G. & Kuznetsov E.N. (1999) PSIC: Profile extraction from sequence alignments with position-specific counts of independent observations. Protein Eng. 12, 387–394.
3. Altschul S.F., Madden T.L., Schäffer A.A., Zhang J., Zhang Z., Miller W. & L ipman D.J. (1997) Gapped BLAST and PSI-BLAST: A new generation of database programs. Nucleic Acids Res. 25, 3389–3402.
4. Sadreyev R. & Grishin N. (2003) COMPASS: A tool for comparison of multiple protein alignments with assessment of statistical significance. J. Mol. Biol. 326, 317–336.
5. Wang G. & Dunbrack R.L. Jr. (2004) Scoring profile-to-profile sequence alignments. Protein Sci. 13, 1612–1626.
6. Noguchi T. & Akiyama Y. (2003) PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB) in 2003. Nucleic Acids Res. 31(1), 492-493.
7. Fiser A. & Sali A. (2003) Modeller: generation and refinement of homology-based protein structure models. Meth. Enzymol. 374, 461-91.

Chen-Tan-Kihara-QA - 653 models for 97 3D/ 34 QA targets

Quality assessment using the diversity of suboptimal alignments

H. Chen[1], Y.F. Yang[1] and D. Kihara[1, 2]
[1] – Dept. of Biological Sciences,[2]- Dept. of Computer Science, Purdue University, West Lafayette, IN, USA
dkihara@purdue.edu

Quality assessment of the structure prediction is crucial for its practical use and this area is not fully developed[1-4]. Our previous work shows the diversity of suboptimal alignments is a good indicator for prediction quality in global or residue level. Based on this strategy, we implement our quality assessment program, a variation of our threading program, SUBWAy, for CASP7.

Our structure prediction program, SUBWAy, can generate a series of suboptimal alignments between the target and the template. Every suboptimal

alignment is denoted as a pathway in the DP matrix plot. The diversity of the assessed alignments can be defined as the average deviation of all suboptimal alignments to the query alignment. Following this concept, a quantitative diversity was assigned to every residue pair in the query alignment (local level) and also to the whole alignment (global level).

Our previous work shows that the local diversity strongly correlates with the distance between the Cα atom in the predicted model and the corresponding atom in the experimental model. The correlation is linearly regressed into:

$$\log(Distance) = 0.3625 * \log(\text{Diversity}) + 1.5672$$

The error estimate on per-residue basis in the QMODE II is calculated from this formula. The global model quality score in the QMODE II is calculated by the following formula:

$$\text{Quality score} = \frac{10}{10 + \text{global diversity}}$$

Our submitted files of quality assessment follow the QMODE II format. Since our method is based on the alignments, in this CASP we only submitted quality assessment for the predictions which provide the predicted alignment files. However, it is very easy to transfer the 3D coordinate model to the structural alignment and predict the quality based on this alignment[5, 6], which will be implemented in next CASP.

1. Mevissen H.T. & Vingron M. (1996) Quantifying the local reliability of a sequence alignment. Protein Eng. 9(2), 127-132.
2. Tress M.L., Jones D. & Valencia A. (2003) Predicting reliable regions in protein alignments from sequence profiles. J. Mol. Biol. 330, 705-718.
3. Cline M., Hughey R. & Karplus K. (2002) Predicting reliable regions in protein sequence alignments. Bioinformatics. 18(2), 306-314.
4. Schlosshauer M. & Ohlsson M. (2002) A novel approach to local reliability of sequence alignments. Bioinformatics. 18(6), 847-854.
5. Shindyalov I.N. and Bourne P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng. 11(9), 739-47.
6. Zemla A. (2003) LGA: A method for finding 3D similarities in protein structures. Nucleic Acids Res. 31, 3370–3374.

CHIMERA - 542 models for 100 3D targets

Protein Structure Prediction using SKE-CHIMERA

M. Takeda-Shitaka, G. Terashi, K. Kanou, D. Takaya, K. Ohta, A.Hosoi, M. Iwadate and H. Umeyama

School of Pharmacy, Kitasato University
shitakam@pharm.kitasato-u.ac.jp

Protein structure prediction using SKE-CHIMERA

In CASP6, we developed SKE-CHIMERA, a web-user interface system for protein structure prediction, through which a lot of data we prepare can be analyzed, and homology modeling is easily carried out with human intervention at necessary stages[1]. Although *CHIMERA-group* succeeded in CASP6, further improvements were required in our method. Therefore, we improved SKE-CHIMERA in CASP7 by preparing much more data for modeling and automating many steps. One of the major improvements is the development of a model evaluation method called CIRCLE (see the abstract of *CIRCLE-group* in the CASP7 Abstracts). CIRCLE score using the 3D1D scoring functions is useful when we select the best model among the models that we constructed for each target.

Side chain refinement targets

In the case of side chain refinement targets, model structures constructed by FAMS[2] were refined by energy minimization and molecular dynamics simulation. After refinement, correct hydrogen bonds were added and the short contacts between atoms were removed. The main-chain conformations constructed by FAMS were not changed largely after refinement.

Multimer prediction targets

We predicted multimer targets using our new program FAMS Complex[3], a fully automated homology modeling system protein complex structures consisting of two or more molecules. FAMS Complex requires only sequences and alignments of the target protein as input and constructs all molecules simultaneously and automatically. FAMS Complex is not docking software that attempts to find the best matching between separate molecules, but is homology modeling software for multi-chain proteins.

Results

The experimental structures of 80 targets have been released as of October 3, 2006. Therefore, we calculated the total score of GDT_TS of every server group including *CHIMERA-group* by simply summing up the GDT_TSs of 80 targets to evaluate the main-chain structures. Moreover, side-chain conformations were evaluated by comparing the side-chain $\chi 1$ torsional angles with those in the native structures for the residues within 3.5 in the MaxSub structure alignment. Side-chain conformations were considered correct if $\chi 1$ were within 40° of the experimental structure values.

GDT_TS	$\chi 1$

Rank	Score (sum)	Server name	Rank	Score (sum)	Server name
1	4908.19	Zhang-Server	1	5329	CHIMERA
2	4811.89	CHIMERA	2	5196	ROBETTA
3	4617.83	Pmodeller6	3	5157	Pmodeller6
4	4604.31	HHpred2	4	5007	FAMSD
5	4574.06	CIRCLE	5	4999	Pcons6
6	4561.34	ROBETTA	6	4952	FAMS
7	4539.88	Pcons6	7	4889	Zhang-Server
8	4539.05	HHpred3	8	4870	CIRCLE

1. Takeda-Shitaka M., Terashi G., Takaya D., Kanou K., Iwadate M. & Umeyama H. (2005) Protein structure prediction in CASP6 using CHIMERA and FAMS. Proteins, Suppl 7, 122-127.
2. Ogata K. and Umeyama H. (2000). An automatic homology modeling method consisting of database searches and simulated annealing. J. Mol. Graphics Mod. 18, 258-272.
3. Takeda-Shitaka M., Terashi G., Chiba C., Takaya D. & Umeyama H. (2006) FAMS Complex: a fully automated homology modeling system for protein complex structures. Medicinal Chemistry, 2, 191-201.

CIRCLE - 500 models for 100 3D targets

CIRCLE: Full automated homology-modeling server using the 3D1D scoring functions

G. Terashi, M. Takeda-Shitaka, K. Kanou, M. Iwadate,
D. Takaya, A. Hosoi, K. Ohta and H. Umeyama

Department of Biomolecular Design, School of Pharmacy, Kitasato University
terashig@pharm.kitasato-u.ac.jp

We have developed CIRCLE server focused on scoring function which evaluates model quality for CASP7. In this server, the scheme is based on searching for the best models from many models, using the 3D1D scoring functions without alignment score, biological information and consensus scoring function such as 3d-jury. A new scoring function refined by CASP6 models was applied to select the models. In the following, we describe the scheme and scoring function.

Method Description

1. Collect structure models and alignments: In the first, the CIRCLE server submitted the target sequence to other alignments or modeling servers (FAMS, FAMSD, FUNCTION, ROBETTA-only alignments, SP3, SPARKS2 and GenTHREADER), and collected results of these servers automatically. For generating refined models from alignments and models, which were collected from other server results, CIRCLE server used our homology modeling program fams[1]. The fams refined side chain conformations, main chain clashes and main chain breaks. In this step, the refined models (100-120 models) were collected to "Structure pool".

2. Predict target difficulty: For predicting the target difficulty, we used Support Vector Machine (SVM) as the classification tool. The training data set was CASP6 targets. The accuracy of this prediction was 85% in CASP6 targets. This predicted difficulty was used in the next evaluation step.

3. Evaluate all models: The all models in "Structure pool" were evaluated by either the scoring functions for CM or FR, NF, which depends on target difficulty as follow.

$$TotalScore = \begin{cases} 0.35 \times SSscore + 3D1Dscore_{CM} & CM \\ 0.75 \times SSscore + 3D1Dscore_{FRNF} & FR\ or\ NF \end{cases} \quad (1)$$

$$SSscore = \sum f(SS_{PREDICTED}, SS_{MODEL}, confidence) \quad (2)$$

SSscore represents the measure of secondary structure similarity (like Q3 value), calculated by comparing secondary structure of model and the result of

PSIPRED[2]. $SS_{PREDICTED}$ represents the secondary structure predicted by PSIPRED. SS_{MODEL} is the secondary structure of model. "*confidence*" is the confidence of prediction, taken from PSIPRED output. $3D1Dscore_{CM}$ and $3D1Dscore_{FRNF}$ are scoring function to evaluate side chain environments. These functions were refined by CASP6 models and difficulties of targets. The 3D1D score is calculated by 3 parameters (fraction of buried area, fraction of polar area, Secondary structure). As shown in function (1),(2), SSscore is given more weight in difficult targets (FR, NF) than easy targets (CM).

Assessment site		category	rank
CAFASP5	MaxSub	ALL	5
	MaxSubDom	ALL	4
TM-Score		ALL	4
Robetta	First_GDT_MM	CM_easy	4
		CM_hard	11
		FR_H	16
		FR_A-NF	9
	First_Z-score	CM_easy	4
		CM_hard	9
		FR_H	18
		FR_A-NF	9
SBC		ALL	5
		EASY	4
		HARD	6

Conclusion

Now (in 9.29.2006) various assessment sites are opened. We summarized ranking of CIRCLE server in 68 CASP servers (remove virtual team and human predictor) in the following table. As shown in this table, the scoring function of CIRCLE server did good selection especially in easy target.

1. Ogata K. and Umeyama H. (2000) An automatic homology modeling method consisting of database searches and simulated annealing. J. Mol. Graphics Mod. 18, 258-272.
2. Jones D.T (1999) Protein secondary structure prediction based on position-specific scoring matrices J Mol Biol/J Mol Biol 292, 195-202.

CIRCLE-FAMS - 496 models for 100 3D targets

Selection from all the server models using original 3D 1D program -"CIRCLE"

G. Terashi, A. Hosoi, M. Takeda-Shitaka, K. Kanou, M. Iwadate, D. Takaya, K. Ohta and H. Umeyama

Department of Biomolecular Design, School of Pharmacy, Kitasato University

terashig@pharm.kitasato-u.ac.jp

The CIRCLE-FAMS team is a meta-selector. In this team, we have selected five models from all the server models (from TS1 to TS5) by using our CIRCLE[1] team method. We describe the details of "CIRCLE-FAMS" as follows.

Collecting server models

All the server models were taken from CASP7 home page.

Refinement of server models

These models include tertiary structures (TS) and alignments (AL), TS models were refined and model structures were generated from alignments by using FAMS[2]. Side-chain conformation, which was refined, was necessary for the evaluation of CIRCLE. These refinement methods were same as our fams-ace team[3].

Evaluation of refinement models

All the refined models were evaluated by CIRCLE. We predicted category (CM or FR) of the target difficulty by Support Vector Machine program[4], we used two kinds of evaluation methods which were the same as our CIRCLE server team. Evaluated models were sorted by 3D1Dscore and 5 high-ranking models with no wrong warning from CASP7 were submitted.

Results

In order to examine the ability of CIRCLE method, we calculated GDT_TS and $\chi 1$[5] angle originally (in 2006/10/3). In the calculation of $\chi 1$ angle, "correct" side chain residue is within 3.5 in the MaxSub superposition and within 40° from native structure. The next table shows the ranking of server teams and our meta-selector team "CIRCLE-FAMS" by using these scores. Category of targets was predicted by our Support Vector Machine program.

GDT_TS			χ1		
Rank	Score (sum)	Server name	Rank	Score (sum)	Server name
1	4908.19	Zhang-Server	1	5326	CIRCLE-FAMS
2	4753.36	CIRCLE-FAMS	2	5196	ROBETTA
3	4617.83	Pmodeller6	3	5157	Pmodeller6
4	4604.31	HHpred2	4	5007	FAMSD
5	4574.06	CIRCLE	5	4999	Pcons6
6	4561.34	ROBETTA	6	4952	FAMS
7	4539.88	Pcons6	7	4889	Zhang-Server
8	4539.05	HHpred3	8	4870	CIRCLE

GDT_TS CM			χ1 CM		
Rank	Score (sum)	Server name	Rank	Score (sum)	Server name
1	3812.47	Zhang-Server	1	4754	CIRCLE-FAMS
2	3722.71	CIRCLE-FAMS	2	4620	ROBETTA
3	3657.96	CIRCLE	3	4559	Pmodeller6
4	3649.04	UNI-EID_expm	4	4539	FAMSD

The above results were obtained by good estimation of the side chain of our original 3D1D CIRCLE method.

1. See "CIRCLE: Full automated homology-modeling server using the 3D1D scoring functions in CASP7" item in this book.
2. Ogata K. and Umeyama H. (2000) An automatic homology modeling method consisting of database searches and simulated annealing. J. Mol. Graphics Mod. 18, 258-272.
3. See "fams-ace: Model selection from server results using original threading program and consensus in CASP7" item in this book.
4. SmartLab, http://www.smartlab.dibe.unige.it/
5. Daniel Fischer, Arne Elofsson, Leszek Rychlewski, Florencio Pazos, Alfonso Valencia, Burkhard Rost, Angel R. Ortiz, and Roland L. Dunbrack, Jr. (2001) Proteins 5 171–183

CIRCLE-QA - 100 models for 100 QA targets

CIRCLE for quality assessment in CASP7

D. Takaya, G. Terashi M. Takeda-Shitaka, K. Kanou, M. Iwadate, A. Hosoi, K. Ohta and H. Umeyama

Department of Biomolecular Design, School of Pharmacy,Kitasato University
p99150@st.pharm.kitasato-u.ac.jp

We have developed CIRCLE[1] since previous CASP because we didn't have high-precision scoring function for tertiary structure. For participation in quality assessment (QA) category of CASP7, CIRCLE-QA aims for ranking server models (TS+AL) by relative score which was proposed by CASP7 organizers. Relative score was calculated based on CIRCLE.

Collecting server models

Server models were obtained from CASP7 home page http://www2.predictioncenter.org/index_serv.html.

Refinement of server models

These models include tertiary structure (TS) and alignment (AL), and therefore these were refined or changed to tertiary structure by FAMS[2]. If it was AL format, a model was built based on this alignment. If it was TS format, a model is refined by FAMS. We used all the server models as its template because these models include CA model or having lacking residue. Moreover, our CIRCLE 3D1D method needs side chain coordinates [1].

Ranking refined models.

CILCLE score corresponding to each predicted difficulty (i.e. CM or FR NF) [3] was calculated for above refined models. We ranked the order using this score. According to the QA rule in CASP7, this score was converted into relative score based on simple rule. The maximum score is 1.0, and the minimum is 0.0. We have adopted the interpretation that model having score of 1.0 is not native structure but best model in each target.

Result

tid	col.	tid	col.	tid	col.	tid	col	tid	col
T0283	0.492	T0303	0.872	T0321	0.368	T0341	0.873	T0363	0.806
T0288	0.91	T0304	0.545	T0322	0.882	T0342	0.773	T0364	0.935
T0289	0.819	T0305	0.931	T0323	0.702	T0345	0.893	T0366	0.852
T0290	0.927	T0306	0.193	T0324	0.858	T0346	0.93	T0367	0.78
T0291	0.831	T0307	0.684	T0325	0.679	T0347	0.582	T0368	0.703
T0292	0.866	T0308	0.78	T0326	0.911	T0348	0.471	T0369	0.44
T0293	0.785	T0309	0.244	T0327	0.794	T0349	0.639	T0370	0.918
T0294	0.887	T0310	0.863	T0328	0.902	T0350	0.685	T0371	0.8
T0295	0.906	T0311	0.727	T0329	0.853	T0351	0.523	T0372	0.707
T0296	0.618	T0312	0.573	T0330	0.892	T0353	0.659	T0373	0.77
T0297	0.69	T0313	0.859	T0331	0.82	T0354	0.528	T0374	0.856
T0298	0.853	T0314	0.377	T0332	0.746	T0357	0.584	T0375	0.826
T0299	0.349	T0315	0.903	T0335	0.619	T0358	0.608	T0376	0.866
T0300	0.516	T0316	0.489	T0338	0.789	T0359	0.695	T0380	0.881
T0301	0.725	T0317	0.937	T0339	0.843	T0361	0.298	T0383	0.795
T0302	0.721	T0318	0.806	T0340	0.92	T0362	0.876	T0384	0.916
								T0385	0.843

This table shows the correlation coefficient in each target between the GDT_TS value and our CIRCLE score. Predicted CM targets(Bold in table) are T0288, T0290, T0291, T0292, T0294, T0295, T0298, T0302, T0303, T0305, T0308, T0310, T0313, T0315, T0317, T0318, T0324, T0326, T0328, T0332, T0338, T0339, T0340, T0341, T0345, T0346, T0359, T0362, T0366, T0371, T0375, T0376 and T0384. Predicted CM targets seem to be high correlation coefficient. Accordingly our CIRCLE is useful in determining the order of modeling quality.

1. See “CIRCLE: Full automated homology-modeling server using the 3D1D scoring functions” item in this book.
2. Ogata K. and Umeyama H. (2000) An automatic homology modeling method consisting of database searches and simulated annealing. J. Mol. Graphics Mod. 18 258-272.
3. See “FAMSD : Homology modeling server providing side chain models with high accuracy” item in this book

CPHmodels - 49 models for 49 3D targets

CPHmodels

Ole Lund, Claus Lundegaard, Morten Nielsen

Center for Biological Sequence Analysis, BioCentrum, Building208. Technical University of Denmark. DK-2800 Lyngby. Denmark.

www.cbs.dtu.dk/services/CPHmodels

Summary

CPHmodels is a server for fold recognition/ homology modeling, in which a large sequence database is iteratively searched to construct a sequence profile until a template can be found in a database of proteins with known structure. The method differs from the PDB-BLAST method in that a sequence profile is only made if a template is not readily found in a database of known structures. A sequence profile is made for the template, using the same number of PSI-BLAST iterations that were used to identify it. Query and template sequences are subsequently aligned using a score based on profile-profile comparisons.

The method is unchanged since 2002, except that the databases are updated each week In CASP5 the alignment score was modified so as to ensure that unreliable parts of the alignment are discarded. The average root mean square deviation (RMSD) for the models which were solved before the CASP5 meeting was 2.3Å. In CASP7 we did not make this modification.

The server is fast and easy to use. We plan to use it in combination with other tools to visualize sequence features and build it together with other prediction servers, such as epitope prediction servers.

Template identification

The program blastpgp [1] was used to search the databases. In order to find a template, the query sequence was run against the pdb database. If a template could not be found with an E value of less than 0.05 the sequence was run against sp, and a binary checkpoint file was saved as well as the position specific scoring matrix in ASCII format. The checkpoint file was used to restart a blastpgp search of the query sequence against the pdb database. The procedure of iteratively using the sp database to generate a profile that in turn is

used to search the pdb database was continued until a template was found with a E value of less than 0.05 or a total number of five iterations against the pdb database had been performed.

Alignment

If a template was identified, we attempted to improve the alignment by performing a profile-profile alignment. In order to make a sequence profile for the template sequence we ran the template sequence the same number of iterations as the query sequence against the sp database and saved the scoring matrix in ASCII format. If no sequence profile was generated for either the query or the template sequence, it was constructed from a blosum62 matrix [2]. A scoring matrix Sij was constructed based on the two profiles.

$$Sij = (QPi(TAj)+TPj(QAi))/2-k$$

Where QPi(TAj) is the score of residue j in the template sequence with the profile at position i in the query sequence, and TPj(QAi) is the score of residue i in the query sequence with the profile at position j in the template sequence. These two scores were averaged and k can be subtracted to reduce the lengths of the alignments and make them more accurate. In the online version of CPHmodels which participated in CASP7 k is set to zero.

Modeling

The corresponding atoms derived from the alignment were extracted from the template file and used as starting point for homology modeling. Missing atoms were added using the segmod program [5], and structures were refined using the encad program [6], both from the GeneMinepackage (www.bioinformatics.ucla.edu/genemine/).

1. Altschul S.F. et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389-3402.
2. Henikoff S., Henikoff J.G. (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A. 89: 10915-10919.
3. McLachlan A .D. (1982) Rapid Comparison of Protein Structres. Acta Cryst. A38: 871-873
4. Shindyalov I.N., Bourne P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng. 11: 739-47.
5. Levitt M. (1992) Accurate modeling of protein conformation by automatic segment matching. J. Mol. Biol. 226: 507-533
6. Levitt M., Hirshberg M., Sharon R. and Daggett V. (1995) Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. Computer Physics Comm. 91: 215-231.

CRACOW.PL - 58 models for 52 3D targets

Simulation of protein folding process rather then protein structure prediction

Irena Roterman[1,2], Michal Brylinski[1,3] , Marek Kochanczyk[1,2]

[1]Department of Bioinformatics and Telemedicine - Collegium Medicum – Jagiellonian University, 31-501 Krakow, Kopernika 17 POLAND

[2]Faculty of Physics, Astronomy and Applied Informatics, Jagiellonian University, 30-059 Krakow, Reymonta 4, POLAND

[3]Faculty of Chemistry, Jagiellonian University, 30-060 Krakow, Ingardena 3, POLAND

The procedure for *in silico* protein folding simulation was applied to construct the structures of target proteins. Two steps process was applied and represented by early- (ES) and late-stage (LS) intermediates. The creation of ES structure is assumed to be determined solely by the backbone conformation [Roterman (1995) *J. Theor Biol* 177, 283-288, Roterman (1995) *Biochimie* 77, 204-216]. The limited sub-space (part of Ramachandran map) distinguished according to low-energy mutual orientation between sequential peptide bond planes appeared to satisfy also the condition of balanced amount of information carried by amino acid in the sequence with the amount necessary to select particular structure of early intermediate [Jurkowski et al. (2004) *Proteins: Struct Func Bioinform.* 55, 115-127]. The contingency table representing the relation between sequences (tetrapeptide unit) and their structures (seven of them distinguished in the limited conformational sub-space) [Jurkowski et al. (2004) *J. Biomol. Struct Dynam.* 22, 149-157, Brylinski et al. (2004) *Bioinformatics* 20, 199-205] created on the basis of complete PDB structures applied to the sequence of the protein of unknown structure allows creation of ES structure which is treated as starting one for the LS step of procedure [Brylinski et al. (2005) *J. Biomed Biotechnol.* 2, 65-79, Meus et al. (2006) *Med Sci Monit* 12, BR208-214, Brylinski et al. (2004) *In Silico Biology* 4, 0022].

The side chain-side chain interaction introduced as the driving force for LS folding step and calculated according to the traditional non-bonding interaction is extended by the hydrophobic interaction. Its presence is expressed by the external force field of hydrophobic character in form of three-dimensional Gauss function (“fuzzy oil drop”) [Konieczny et al. (2006) *In Silico Biology* (2006) 6, 15-22]. The conformational changes of folding molecule decreasing the difference between idealized and observed hydrophobicity density distribution are accepted. The hydrophobic core in a central part of “fuzzy-oil-drop” is created in consequence of this procedure. Surface of the molecule gets covered by hydrophilic residues. The starting size of “oil drop” is determined by the size of ES structure. Its size gets decreased step-wise reaching the size characteristic for the molecule of particular polypeptide chain length. Every

step of hydrophobicity oriented optimization is followed by the traditional non-bonding energy optimization oriented procedure (ECEPP force field) to eliminate the possible overlaps. The procedure stops when non-bonding energy convergence criterion is reached and the size of drop is similar to the expected one for protein of particular polypeptide length [Brylinski et al. (2006) *J. Biomol Struct Dynam*. 23, 519-527, Brylinski et al. (2006) *Biochimie* 88, 1229-1239, Brylinski et al. *Comp Biol Chem* 30, 255-267].

The ES step of folding process was tested in CASP6 and the LS step was applied in CASP7 [Konieczny et al. (2006) *In Silico Biology* (2006) 6, 15-22].

The top results in the structure prediction are not expected by the group. The "fuzzy oil drop" model produces the very well soluble protein covered by hydrophilic residues with no biological activity. The "fuzzy oil drop" model applied to crystal structures of proteins reveals the significant discrepancy between idealized ("fuzzy oil drop" model) and empirically observed distribution of hydrophobicity density localized exactly in the area of substrate or ligand binding. This observation suggests important role of "ligand" or "substrate-like" molecule in folding process (what was shown in ribonuclease folding simulation [Brylinski et al. (2006) *Comp Biol Chem* 30, 255-267]. As long as the active participation of ligand or ligand-like molecule is not taken into account in folding simulation, the structure prediction applying the "fuzzy-oil-drop" model can not be successful. Although (as shown in the TA03354_69_121 target of CASP6 [Konieczny et al. (2006) *In Silico Biology* 6, 15-22] in some small proteins the possible binding cavity can appear without the ligand present.

The conclusion is that the active presence of ligand or substrate-like molecule is necessary during folding process at least in folding process *in silico.*

Dill-ZAP - 30 models for 6 DR targets

Physics-Based Protein Folding by Zipping and Assembly

M. S. Shell[1], S. B. Ozkan[1], V. Voelz[2], G. A. Wu[1], V. Coutsias[3], J. Chodera[2], R. Ritterson[2], S. Cordes, K. Dill[1]

[1] – *Department of Pharmaceutical Chemistry, UC San Francisco,*
[2] - *Graduate Group in Biophysics, UC San Francisco,*
[3] - *Department of Mathematics and Statistics, Univ. of New Mexico*
shell@maxwell.compbio.ucsf.edu

Our goal in CASP was to be as physical as possible. Our scoring function is just a physics-based force field (Amber 96 + GB/SA solvent). We do not use protein database information, such as secondary structure preferences or PDB-based potentials, or low-resolution starting models. For sampling, we aim to mimic physical folding routes: (1) We do local searching by replica exchange molecular dynamics (REMD), to ensure proper Boltzmann populations. (2) Our global searching involves mechanistic folding routes that we learn on the fly. Conformational sampling occurs along zipping and assembly routes (our algorithm is called ZAM – Zipping & Assembly Method). As an offshoot, this method also makes predictions about: (1) the physical folding routes (for four non-CASP PDB proteins, we predict roughly correct Phi distributions) and (2) protein stability (however, the current force field gives ion-pairs that are too strong).

In the Zipping & Assembly mechanism, an unfolded chain first explores locally favorable structures at multiple independent positions along the chain. These local structures tend to have hydrophobic contacts and contain small α-helical or β-turn structures. While only transiently stable on their own, such local structures can then either: (a) grow (which we call zipping) by recruiting neighboring amino acids in the sequence to form additional contacts, or (b) come together as units (assembly). In these ways, the protein chain grows increasingly ordered and native-like.

More specifically, ZAM works as follows:

1. The full protein is parsed into overlapping 8-mer fragments spaced every 3 residues apart. Each such fragment begins in the extended state, and is then energy-minimized with AMBER ff96 + the Generalized Born implicit solvation model of Onufriev, Bashford, and Case, followed by 5 ns of REMD, in the absence of the rest of the chain.

2. We retain those 8-mers that satisfy either of two criteria: (a) either they have persistent structure (see below), or those 8-mers can recruit additional local residues to grow more structure cooperatively (i.e., with non-additive free energies), as determined by a look-ahead analysis (PUNCH). New chain is then added to those 8-mers, to grow them into 12mers, followed by REMD for another 5 ns. The process is repeated to reach partially structured 16-mers.

3. Stable contacts are identified within each 16-mer fragment using the potential of mean force (PMF) vs. distance for all possible hydrophobic residue pairs in each fragment, computed by weighted histogram analysis (WHAM). We take the residue pairs for which the PMFs show a pronounced minimum in free energy at a distance less than 8.0 Å as favorable and stable (i.e. sampled at least 50% of the time). Any fragment having mutually exclusive (i.e., "competing")

stable contacts is split into separate ensembles in which that fragment has either of the two possible contacts.

4. To enforce any particular emerging folding route, a stable contact is locked into place by imposing a harmonic restraint between residue centroids with a force constant of 0.5 kcal / (mol Å^2). Fragments are then grown by adding new residues at each terminus, followed by 5 ns REMD simulations. Thus, most of the new sampling focuses on the newly added residues, largely avoiding re-sampling the existing structure. Steps 3 and 4 are iterated until fragments cannot be grown further, and until no new contacts are persistent.

5. When fragments cannot zip further, assembly of existing fragments is attempted, in two steps: (a) We generate a distribution of rigid body arrangements of the two structured fragments (PHAT – packing by hydrophobic alignment tool). (b) The loops are connected and sampled by a fast analytical robotics-based method (called SPLAT – Sampled Protein Loop Assembly Tool). The assembled structures are clustered and ranked by hydrophobic radius of gyration, and the top-ranked structures are used as initial conformations for another round of REMD simulations. This gives a fast way to sample possible topological assemblies.

6. Our sampling of this physical force field remains severely limited, so we cannot directly compute the relative free energies of the various possible final states. Instead, we filtered our final structures, keeping only those that have small hydrophobic radius of gyration.

DISpro (server, disorder) - 100 models for 100 DR targets

Protein Disordered Region Prediction Using DISpro

Jianlin Cheng, Mike Sweredoski, and Pierre Baldi
Institute for Genomics and Bioinformatics, School of Information and Computer Science
University of California Irvine, CA 92697

Intrinsically disordered regions in proteins are relatively frequent and important for our understanding of molecular recognition and assembly, and protein structure and function. Our *ab initio* predictor of disordered regions called DISpro participated in CASP7. DISpro [1] uses evolutionary information in the form of profiles, predicted secondary structure and relative solvent accessibility, and ensembles of 1D-recursive neural networks to predict disordered region.

1. Cheng J., Sweredoski M., and Baldi P. (2005) Accurate Prediction of Protein Disordered Regions by Mining Protein Structure Data, Data Mining and Knowledge Discovery, vol. 11, no. 3, pp. 213-222.

Distill - 800 models for 100 3D/100 DP/100 DR/100RR targets
Distill_human - 800 models for 100 3D/100 DP/100 DR/100RR targets

Draft protein structures by machine learning

Davide Baù[1], Alberto J. Martin[1], Catherine Mooney[1], Alessandro Vullo[1], Ian Walsh[1], Silvio Tosatto[2], Gianluca Pollastri[1*]
School of Computer Science and Informatics,
University College Dublin, Ireland
CRIBI, Università di Padova, Italy
** gianluca.pollastri@ucd.ie*

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

Secondary structure is predicted by Porter[1], relative solvent accessibility by PaleAle[2], contact density by BrownAle[3], residue contact and distance maps by XXStout[3]. Residue contact maps submitted to CASP (8Å) are obtained by XXStout, and are not directly used to predict 3D coordinates. 4-class distance map predictions by an architecture identical to XXStout's are adopted instead. All structural feature predictors are based on single- or dual-layer Recursive Neural Network architectures for Directed Acyclic Graphs (DAG RNNs)[4]. One-dimensional feature predictors (i.e. those mapping the primary sequence into a sequence of the same length) are based on 1D DAG RNNs, while contact and distance map predictors are based on 2D DAG RNNs. Secondary structure, solvent accessibility and distance map predictors are provided structural information about PDB templates as a further input, when templates are available. Templates are identified as follows: 2 rounds of PSI-BLAST are run against UniProt; the resulting PSSM, plus predictions of structural motifs by

Porter+[5], are aligned locally against all the sequences and corresponding structural motifs in the PDB. Because of a glitch in the updating procedure, at CASP we used an outdated PDB (March 2005), resulting in suboptimal predictions for numerous targets.

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

Initial structures are generated, in which helices predicted by Porter are modelled, consecutive Cα atoms are set at a realistic distance (~3.8Å), and virtual Cα angles are restricted to the 90º-180º interval.

A search from these initial structures is performed by introducing perturbations in them. Helices are treated as rigid "rods" and their core Cαs are never moved on their own. The search is carried out by simulated annealing with a linear schedule for the temperature. 5,000 moves of every non-helical Cα and helical termini are attempted for each search. 50 searches are run for each protein structure.

Finally, the structures obtained are ranked. In the *ab initio* case we rank the structures by a neural network trained to map a number of characteristics (enforcement of of predicted constraints, secondary structure composition, compaction, etc.) of each structure into its quality, measured as its TM score against the correct structure. In the case templates from the PDB are available, similarity to the templates is used as further information for ranking.

We also submitted predictions of protein domains and protein disorder by predictors that are not integrated in Distill's pipeline. The predictor or protein domains (Shandy) has three stages: one in which proteins are classified as most likely single-domain vs. possibly multi-domain (currently implemented as a hard threshold of 180 residues); a second stage (a 1D DAG Recurrent Neural Network) in which residues in the latter proteins are marked as domain boundary vs. intra-domain; a third stage in which the previous predictions are smoothed and the location of domain boundaries is decided. Disorder is predicted by Spritz[6], a combination of experts implemented by kernel machines.

Distill_human is the same, fully automated predictor as Distill (see abstract). Given the looser deadlines for human submission, all predictions by Distill_human are based on an improved fold recognition component which was introduced into Distill after the first 10 targets. All the following submissions are identical to Distill.

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

DISTILLFM - 68 models for 68 RR targets

A Filtering Approach for Improved Modeling of Predicted Contact Maps

Alberto J. Martin, Alessandro Vullo and Gianluca Pollastri

School of Computer Science and Informatics
University College Dublin
Belfield, Dublin 4, Ireland
{albertoj,alessandro.vullo,gianluca.pollastri}@ucd.ie

For CASP7 we predicted contact maps as follows: contact information is first predicted by our ab initio algorithm (XXStout [1]); an ensemble of multi-layered perceptrons filters the initial predictions. The mapping is implemented by providing the filtering ensemble with information about physical realisability, violations of basic principles and long range contact information observed in the predicted contact maps.

XXStout uses information from multiple sequences alignment profiles, predicted secondary structure, solvent accessibility and contact density. XXStout's predictions contain errors, mainly because they are local: the patterns of contacts between secondary structure elements are often shaped differently from those found in real contact maps; some amino acids are in

contact with too many other amino acids or with too few, making predicted maps not physically realisable.

The filtering stage is implemented by training two different sets of multilayered perceptrons to filter different positions of the contact map; one for positions close to the diagonal, one for positions far from the diagonal, since the rules governing contact probability are likely to be different for the former case (mainly made by backbone atoms, reflecting secondary structure) than for the latter (where contacts mainly occur between the side chains of amino acids placed in different secondary structure elements). For each pair of residues (i,j), the input features of both learners are the following: a square window of predicted contact probabilities (by XXStout) centred on (i,j); predicted secondary structure in three states, by Porter [2]; the number of amino acids predicted to be contact with either residue i or j; the number of amino acids predicted to be in contact with both residues; i's and j's contact order [3].

1. Vullo A., Walsh I., and Pollastri G. (2006) A two-stage approach for improved prediction of residue contact maps. BMC Bioinformatics, 7:180.
2. Pollastri G. and McLysaght A. (2005) Porter, a new, accurate server for protein secondary structure prediction. Bioinformatics, 21(8):1719–1720.
3. Plaxco K.W., Simons K.T., and Baker D. (1998) Contact order, transition state placement and the refolding rates of single domain proteins. J. Mol. Biol. 277, 985–994.

dokhlab - 114 models for 26 3D targets

Ab Initio Structure Prediction by Rapid Sampling of Protein Conformational Space

F. Ding, A.W. Serohijos, S. Sharma, H. Nie, S. Yin, N.V. Dokholyan

Department of Biochemistry and Biophysics, School of Medicine, UNC-CH
dokh@med.unc.edu

One of the primary bottlenecks in the *ab initio* protein structure prediction is the adequate sampling of the vast conformational space on which the native structure of the protein resides. Here, we develop an *ab initio* protein structure methodology that rapidly searches the protein energy landscape for the global energy minimum. We use a dynamic sampling algorithm, discrete molecular dynamics (DMD[1,2]), to rapidly sample the conformational space, and a recently-developed all-atom energy function to guide the DMD simulation toward the low-energy state[3].

First, we reduce the protein conformation available to the target sequence by constructing a linear peptide with secondary structures (helix, beta strand, and random coil) derived from consensus among several secondary structure prediction schemes (e.g., PSIPRED, SAM, PROFSEC). In regions with low consensus secondary structure prediction, we generate different models with either helix, beta strand, or random coil constraints assigned to the low confidence regions.

Second, we fold the linear peptides using replica exchange[4] DMD simulations. Using square-well potentials and collision driven dynamics, DMD is able to rapidly sample the protein conformational space and follow long time scale events of proteins such as folding. In the folding simulation, we employed united atom models of proteins which include all heavy atoms and polar hydrogens. The energy of a protein conformation is computed as a linear sum of the following terms:

$$E = W_{bonded}E_{bonded} + W_{vdw}E_{vdw} + W_{solv}E_{solv} + W_{bb\_hbond}E_{bb\_hbond} + W_{sc\_hbond}E_{sc\_hbond} + W_{bb\_sc\_hbond}E_{bb\_sc\_hbond}$$

Here, E_{bonded} is the bonded energy, E_{vdw} and E_{solv} are the VDW and solvation energies, E_{bb_hbond}, E_{sc_hbond} and $E_{bb\_sc\_hbond}$ are the hydrogen bond energies among backbones, among sidechains, and between backbones and sidechains, respectively. The detailed description of the energy terms and the parametrization is described in Ref.[1]. The secondary structures from the first step are incorporated into the simulations as constraints to quickly bias the initial linear chain into the defined secondary structures. For each secondary structure model as defined in the first step, we perform two rounds of replica exchange simulations. In the first round, we sample the neighborhood of the protein folding transition by replica simulations around the melting transition. We then construct a preliminary decoy set composed of structures with low values of potential energies and radii of gyration. All the decoys generated from different secondary structure schemes are subject to clustering algorithm. The centroids for the top-populated clusters are selected for the next round of simulations. The objective of the first round of simulation is to transform the linear peptide into molten globular state with correct topological wiring. In the second round of folding simulations, each structure in the preliminary decoy set was subjected to replica exchange-DMD with exchange temperatures below the folding transition. This round of folding simulation improves hydrogen bond pattern formations and protein packing density by eliminating void spaces within the structure. We then construct a final decoy set based on the following criteria: low energy, hydrophobic packing, minimal buried charged residues, and minimal internal void spaces.

Lastly, as a final stage refinement, we optimize the side-chain rotamer states of the candidate structure using Monte Carlo-based search for low-energy rotamer states. For a given rotamer state, there are associated dihedral angle variations with their standard deviations tabulated in a rotamer library[5]. A trial rotamer is

rejected or accepted based on the difference of the minimum energy of the trial rotamer after minimization and the original energy.

1. Ding F. & Dokholyan N. V. (2005) Simple but predictive protein models. Trends in Biotechnology 23, 450-455.
2. Zhou Y.Q., Karplus M., Wichert J.M. & Hall C.K. (1997) Equilibriumthermodynamics of homopolymers and clusters: Molecular dynamics and Monte Carlo simulations of systems with square-well interactions. Journal of Chemical Physics 107, 10691-10798.
3. Ding F. & Dokholyan N. V. (2006) Emergence of Protein Fold Families through Rational Design. PLoS Computational Biology 2, e85.
4. Sugita Y. & Okamoto Y. (1999) Replica-exchange molecular dynamics method for protein folding. Chemical Physics Letters 314, 141-151.
5. Dunbrack J.R. & Cohen, F.E. (1997) Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci 6, 1661-1681.

dokhlab - 114 models for 26 3D targets

Atomic-Resolution Prediction of Protein Structure Using Constrained Replica-Exchange Annealing

F. Ding, A.W. Serohijos, S. Sharma, H. Nie, S. Yin, N.V. Dokholyan

Department of Biochemistry and Biophysics, School of Medicine, UNC-CH
dokh@med.unc.edu

Prediction of tertiary structure of proteins based on comparative modeling is one of the most ubiquitous approaches for protein structure prediction with atomic level accuracy. Often target sequence has a high homology with proteins having experimentally known structure, in which cases a very accurate prediction of the target sequence can be achieved using comparative modeling. However a major bottleneck of homology-based structure prediction techniques is of achieving adequate conformational sampling to find the most stable tertiary structure for a putative secondary structure predicted by homology. Here we present a biophysically-principled approach for predicting the tertiary structure of proteins using comparative modeling followed by replica-exchange simulations to achieve the global energy minima[1]. We exploit the rapid conformational sampling abilities of discrete molecular dynamics (DMD) [2,3] to reach the minimum energy conformation for the protein

In this approach, the 3D-Jury Metaserver[4] (http://bioinfo.pl/meta/) is used to assess a consensus homology prediction of secondary structure for the target sequence. Multiple databases available at the 3D-Jury server are used to generate a consensus prediction - the target sequence is submitted to 3D-Jury meta-server which scores putative structural models based on their similarity to other models. A similarity metric (J-score) is assigned by the 3D-Jury server, corresponding to the number of C-alpha atoms after superposition within 3.5 Å root mean square deviations from the native structure. The most homologous structure for the target sequence, as predicted by the 3D-Jury Metaserver is selected to define the secondary structure of the target.

In the second step, the predicted secondary structure is used to ascribe interaction-constrains between all pairs of heavy-atoms in the structure which define the secondary structure. Using the MEDUSA software[1], a linear-chain model of the peptide is generated along-with the secondary structure constrains and the MEDUSA force-field[1] designed for all-atom DMD simulations of proteins. Next we perform a short-duration, low temperature simulation for relaxing the linear conformation of the protein and generating multiple initial structures to be used in replica-exchange simulation. We then perform multiple replica exchange annealing simulations of this model using these initial conformations, and eight replicas for relaxing the protein structure under the secondary structure constrains. Upon completion of replica-exchange simulations, tentative predictions having lowest energy among all replicas are selected and another run of replica-exchange simulations is performed, starting with these tentative predictions. Upon completion of second round of replica exchange simulations, structures having lowest mean radii are selected (i.e. those structures which form compact topologies) as the putative heavy atom structure of the target protein.

Finally, hydrogen atoms are added to the putative heavy atom structure and the side-chain packing, rotamer states of these structures is then optimized using fixed-backbone redesign module of MEDUSA software[1]. In this step, a Monte Carlo-based search for low-energy rotamer states is performed using the Dunbrack rotamer library[5]: For a given rotamer state, there are associated dihedral angle values with their fluctuations are recorded in a rotamer library. A trial rotamer is rejected or accepted based on the difference of the minimum energy of the trial rotamer after minimization and the original energy.

1. Ding F. & Dokholyan N.V. (2006) Emergence of protein fold families through rational design. PLoS. Comput. Biol. 2, e85.
2. Ding F. & Dokholyan N.V. (2005) Simple but predictive protein models. Trends Biotechnol. 23, 450-455.
3. Ding F., Dokholyan N.V., Buldyrev S.V., Stanley H.E. & Shakhnovich E. (2002) I. Direct molecular dynamics observation of protein folding transition state ensemble. Biophys. J. 83, 3525-3532.
4. Ginalski K., Elofsson A., Fischer D. & Rychlewski L. (2003) 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics. 19, 1015-1018.

5. Dunbrack R.L., Jr. & Cohen F.E. (1997). Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci. 6, 1661-1681.

DomFOLD- 100 models for 100 DP targets

A combined approach to automated protein domain prediction

L.J. McGuffin[1]

1 – The Bioinformatics and Systems Biology Unit, The BioCentre, The University of Reading, Whiteknights,Reading RG6 6AS, UK
l.j.mcguffin@reading.ac.uk

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

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

The second method parses the top alignments from mGenTHREADER. Domain boundaries are assigned by the location of each fold aligned to the target sequence. Where possible, the boundaries of aligned folds with multiple domains are appropriately subdivided using the SCOP domain assignment. The alignment rankings are used to discriminate between conflicting domain assignments, i.e. where domain boundaries of different folds overlap the mGenTHREADER score is used to select the highest ranking domain. Therefore, the overall domain assignment for this method is essentially determined by the top model.

The third method is similar to that above, however data from multiple models are used to determine boundaries. Alignments from the top five nFOLD models are used to provide five alternative domain assignments. The consensus domain assignment is then used to determine overall domain boundaries for this method.

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

1. Marsden R., McGuffin L.J. & Jones D.T. (2002) Rapid protein domain assignment from amino acid sequence using predicted secondary structure. Protein Sci., 11, 2814-2824.
2. Jones D.T., Bryson K., Coleman A., McGuffin L.J., Sadowski M.I., Sodhi J.S. & Ward J.J. (2005) Prediction of novel and analogous folds using fragment assembly and fold recognition. Proteins: Structure, Function and Bioinformatics. 61 (S7), 143-51.
3. Ward J.J., Sodhi J.S., McGuffin L.J., Buxton B.F. & Jones D.T. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J. Mol. Biol., 337, 635-645.
4. Jones D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol. 292, 195-202.
5. Kabsch W. & Sander C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 22, 2577-637.
6. Murzin A.G., Brenner S.E., Hubbard T. & Chothia C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536-540.

Duan_group- 13 models for 7 TR targets

Protein Structure Refinement by Molecular Dynamics Simulation

T. Wang and Y. Duan

Genome center and Bioinformatics
University of California, Davis, CA 95616, USA
duan@ucdavis.edu

We participated in the CASPR experiment and the structure refinement session of the CASP7 experiment. We used Molecular Dynamics (MD) simulations to refine the structures that were predicted in CASP experiments. The simulations were carried out by using the AMBER8 program[1] with the ff03 force field[2] and the Generalized Born (GB) implicit solvent model[3]. To remove serious steric clashes in the starting structures, energy minimization was conducted prior to MD simulations, until energy gradient reached 10 kcal/mol·Å. No cutoff was used in the energy minimization. In MD simulations, the minimized structures were first heated gradually from 10K to 300K in 40 ps and then the temperature was maintained with a temperature-coupling constant of 1.0 ps for 10 ns. A time step of 1 fs was used, and the non-bonded interactions were updated every 25 time steps with a cutoff of 12Å.

The simulation trajectories were analyzed by computing the root-mean-square deviation (RMSD) against the experimental structures and the total energies. Two models were submitted for each of the refinement target: one with the lowest RMSD (or the last snapshot structure in the blind test cases) and the other with the lowest energy. The lowest RMSDs and the lowest energies varied with different targets. Small RMSDs were obtained for the proteins that are monomers in their biological forms, for example, 1.21Å in the case of target TMR04 and 1.88Å in the case of target TMR01. But for the proteins that are dimmers and tetramers, simulations of their monomers generated structures far from the starting structures and large conformational changes were observed because of lack of the stabilization from other monomer counterparts. Other factors that affected the performance of the refinement were the RMSDs of the starting structures and the severity of bad contacts in the starting structures.

1. Case D. A., Cheatham,T. E., Darden T., Gohlke H., Luo R., Merz K. M., Onufriev A., Simmerling C., Wang B. & Woods R. J. (2005) The Amber biomolecular simulation programs. J Comput Chem. 26, 1668-1688.
2. Duan Y., Wu C., Chowdhury S., Lee M. C., Xiong G., Zhang W., Yang R., Cieplak P., Luo R., Lee T., Caldwell J., Wang, J. & Kollman P. (2003) A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. J. Comput. Chem. 24, 1999-2012.
3. Onufriev A., Bashford D. & Case D.A. (2004) Exploring protein native states and large-scale conformational changes with a modified generalized born model. Proteins. 55, 383-394.

EAtorP- 20 models for 20 3D targets

Ab Initio Protein Structure Prediction with a Dipeptide-Assembly Evolutionary Algorithm

A. Bazzoli[1], A.G.B. Tettamanzi[1] and G. Colombo[2]

[1] - Dipartimento di Tecnologie dell'Informazione, Università degli Studi di Milano, via Bramante 65, 26013, Crema, Italy [2] - Istituto di Chimica del Riconoscimento Molecolare, CNR, via Mario Bianco 9, 20131, Milano, Italy

bazzoli@dti.unimi.it

The prediction method is an evolutionary algorithm that identifies the native structure as the global minimum of an energetic fitness function on a discretized conformational space.

A heavy-atom representation of the backbone is used, where bond lengths and angles have ideal values[1] and peptide units are assumed to be planar. Side chains are added to all non-Glycine residues, and are simplified to a single beta carbon that is located along a fixed direction[1] but whose distance from the alpha carbon depends on aminoacid type[2]. The conformational degrees of freedom thus reduce to the mere sequence of PHI and PSI backbone torsional angles. As a further constraint, model conformations are built through dipeptide fragment assembly, where each fragment is represented by a contiguous (PSI, PHI) pair extracted from the PDB_SELECT25 database[3]. The entire collection of extracted pairs has been partitioned by dipeptide aminoacid type into 20x20=400 subsets, so that each dipeptide in the query protein is forced to sample only pairs of its own type. Within each subset, pairs are identified by a numerical index, allowing conformations to be encoded as strings of integer genes.

Each run of the evolutionary algorithm is 1600 generations long and uses a population of 800 conformations. Coarse-grained exploration of the search space is carried out by one-point crossover and single-gene mutation: the former, applied at a 0.7 rate, makes two conformations exchange their C-terminal portion, whose length depends on where the crossover point falls; the latter, applied at a 0.001 rate, blindly replaces one of the conformation's (PSI, PHI) pairs with a random pair from the same subset, producing a rotation of the C-terminal portion that starts at the mutation site. As a special case of single-gene mutation, a fine-tuning operator has also been devised, which allows to explore the neighbourhood of a given conformation by replacing a (PSI, PHI) pair with a similar one. This operator is used in the context of a local-search process, that is applied to each conformation with probability 1.0 and adapts itself to be either an optimization or exploration tool according to the distribution of fitness values among the population [4].

The fitness of a conformation is defined to be a linear combination of three quantities: the first measures steric violations by adding a penalty term for each pair of atoms at a distance less than their summed van der Waals radii; the second is the pairwise contact energy of residues, calculated from a previously-reported contact-energy table[5]; and the third is the radius of gyration of the conformation. The relative weights of these three quantities have been determined by experiments on a training set of 12 proteins[6], which were aimed at finding a general correlation between fitness function and RMSD to the native state. Interestingly, the best correlation, together with a sufficient steric feasibility of conformations, is achieved when the radius of gyration is assigned a weight far greater than the other two.

1. Engh R.A. & Huber R. (1991) Accurate bond and angle parameters for X-ray protein structure refinement. Acta Cryst. A47, 392-400.
2. Feig M., Rotkiewicz P., Kolinski A., Skolnick J. & Brooks C.L. (2000) Accurate reconstruction of all-atom protein representations from side-

chain-based low-resolution models. Proteins: Struct. Funct. Genet. 41, 86-97.
4. Hobohm U., Scharf M., Schneider R. & Sander C. (1992) Selection of a representative set of structures from the Brookhaven Protein Data Bank. Protein Science. 1, 409-417.
5. Krasnogor N. & Smith J. (2000) A memetic algorithm with self-adaptive local search: TSP as a case study. Proceedings of the Genetic and Evolutionary Computation Conference. 987-994. Morgan Kaufmann.
6. Berrera M., Molinari H. & Fogolari F. (2003) Amino acid empirical contact energy definitions for fold recognition in the space of contact maps. BMC Bioinformatics. 4, 8.
7. Gong H., Fleming P.J., & Rose G.D. (2005) Building native protein conformation from highly approximate backbone torsion angles. Proc. Nat. Acad. Sci. 102, 16227-16232.

EBGM-LBT- 40 models for 15 3D targets

Assessing a new approach for protein structure modeling combining structural alphabet local conformation prediction and greedy algorithm for reconstruction.

J. Maupetit[1], F. Guyon[1], J. Martin[1], A.C. Camproux[1], Ph. Derreumaux[2] and P. Tuffery[1]

1 - INSERM U726, 2 - CNRS UPR 9080, Paris France
tuffery@ebgm.jussieu.fr

We have developed a modeling approach that relies on the concept of structural alphabet (SA), i.e. the description of the local structure of proteins using prototype conformations. Here we use a Hidden Markov Model (HMM) derived structural alphabet (HMM-SA) of 27 “letters”1. Each letter of the alphabet describes the conformation of fragments of 4 residues length, consecutive fragments overlap by 3 residues. Each protein can thus be described by a series of SA letters, or trajectory.

Our modeling process relies on three steps: (i) prediction of the local structure of proteins from the amino-acid sequence (ii) search for structural candidates local or global using a similarity search facility based on the alignment of series of letters of the structural alphabet. (iii) fragment assembly.

The prediction of the SA trajectory from the amino-acid sequence results from a learning over 3439 proteins. It can be constrained by the results of secondary structure prediction tools such as PSI-PRED2, or the knowledge of the conformation of regions of the structure. Starting from this predicted description of protein structures, we search for 3D candidate fragments (manuscript in preparation) matching that prediction in the PDB, using classical sequence alignment methods transposed to SA3.

The assembly of the candidate fragments is achieved using a greedy algorithm. Starting from a description of the protein as overlapping fragments (HMM-SA letters), we have recently shown that stochastic greedy algorithm is able to rebuilt protein structures with a satisfying accuracy4, using RMSd or Go potential as objective functions. For CASP, we have used a modified version of the OPEP force field5 to drive the greedy algorithm during the rebuilding process.

For CASP7, we have assessed two different strategies for modeling. In the first, we start from the predicted HMM-SA description conditioned by PSI-PRED profiles, refined by the search for candidate fragments against the PDB. Such strategy was used for the de novo modeling. In the second, used for comparative modeling, we use information from a template as a constraint. Possible templates have been determined using the 3D-Jury6 Meta Server (http://bioinfo.pl/Meta/), a local implementation of PDB-Blast, or our structural similarity search tools. In the aligned regions, we use the SA description of the template. The prediction of loop conformations is constrained by the trajectory of the template on the flanking regions.

A final refinement is performed using Gromacs7, and SABBAC8.

1. Camproux AC., Gautier R., Tuffery P. (2004) A hidden markov model derived structural alphabet for proteins. J Mol Biol. 339, 591-605.
2. Jones D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292: 195-202.
3. Guyon F., Camproux A.C., Hochez J., Tuffer P. (2004) SA-Search: a web tool for protein structure mining based on a Structural Alphabet. Nucleic Acids Res. 32, W545-8.
4. Tuffery P., Derreumaux P. (2005) Dependency between consecutive local conformations helps assemble protein structures from secondary structures using Go potential and greedy algorithm. Proteins. 61, 732-40.
5. Santini S., Wei G., Mousseau N., Derreumaux P. (2003) Exploring the Folding Pathways of Proteins through Energy Landscape Sampling: Application to Alzheimer's β-Amyloid Peptide. Internet Electron. J. Mol. Des., 2, 564-577
6. Ginalski K., Elofsson A., Fischer D., Rychlewski L. (2003) "3D-Jury: a simple approach to improve protein structure predictions." Bioinformatics. 19(8):1015-8.
7. Van Der Spoel D., Lindahl E., Hess B., Groenhof G., Mark A.E., Berendsen H.J. (2005) GROMACS: fast, flexible, and free. J Comput

Chem. 26(16):1701-18.
8. Maupetit J., Gautier R., Tuffery P. (2006) SABBAC: online Structural Alphabet-based protein BackBone reconstruction from Alpha-Carbon trace. Nucleic Acids Res. 34:W147-51.

FAIS - 338 models for 78 3D/100 DR targets

Protein tertiary structure prediction based on contact number prediction

T. Ishida[1], and K. Kinoshita[1]

1 - Human Genome Center, The Institute of Medical Science, The University of Tokyo

t-ishida@hgc.jp

We used a potential based on contact number prediction for both homology and *de novo* prediction. This potential uses the contact number of residues in a protein structure and the absolute contact number of residues predicted from its amino acid sequence using a prediction method based on a support vector regression (SVR) [1]. The contact number of an amino acid residue in a protein structure is defined by the number of residues around a given residue. First, we predicted the contact number of each residue using SVR from Position Specific Score Matrices (PSSMs) in a 15-residue window centered on the target residue. Then, the potential of the protein structure is calculated from the probability distribution of the native contact numbers corresponding to the predicted ones.

To predict protein structures, we first searched templates using PSI-BLAST and FORTE[2] server via the Net. If there were some good templates, we generated 100 models for each template by using Modeller. We selected final models from these predicted models using the above potential based on contact number prediction.

For NF targets, we produced tertiary structure models by using our *de novo* modeling system based on the general fragment assembly method[3]. We searched candidate fragments of each position using the Pearson's correlation coefficient between the PSSMs of a query subsequence and the PSSMs of a target subsequence. Using the fragment libraries, we searched conformational spaces using a potential energy function by simulated annealing method. Our potential energy function includes a term of above potential based on contact number prediction, atom clashes, and hydrogen bonding. We produced about 10,000 models for each target, and selected 5 prediction models by using the potential energy and structural clustering. Finally, sidechain modeling was performed by using SCWRL version 3.0.

1. Ishida T., Nakamura S., Shimizu K. (2006) Potential for assessing quality of protein structure based on contact number prediction. Proteins, 64, 940-947.
2. Tomii K. and Akiyama Y. (2004) FORTE: a profile-profile comparison tool for protein fold recognition. Bioinformatics, 20, 594-595.
3. Simons K.T., Kooperberg C., Huang E. & Baker D. (1997) Assembly of Protein Tertiary Structures from Fragments with Similar Local Sequences using Simulated Annealing and Bayesian Scoring Functions. J. Mol. Biol. 268, 209-225.

FAMS - 500 models for 100 3D targets

Automatic modeling server using homology modeling method and *ab-initio* method

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

[1] - Department of Biomolecular Design
School of Pharmacy, Kitasato University
p01055@st.pharm.kitasato-u.ac.jp

FAMS is the homology modeling server including the *ab-initio* method. Homology modeling method is effective if the homologous proteins of the target are found, but not effective when no homologous proteins are found. In the previous CASP experiment, many models constructed using the homology modeling program "FAMS" [1] were good in the CM category, but not so good in the non-CM, especially NF category. In this time, FAMS server had included *ab-initio* method based on fragment assembly (we call 'TEMPLA'; TEMPlate-Less *Ab-initio*) for NF-targets.

Method Description

1. Constructing structure using TEMPLA

If the alignment score of sparks2[2] is higher than 4.5 and the length of target sequence is less than 150, TEMPLA had been executed with following fragment assembly algorithm.

In the first step, the peptide fragments were generated according to segment-distance calculated by equation as follows [3]:

$$DISTANCE = f(AAfrequencies, SSconfidence)$$

where *AAfrequencies* is the frequencies of amino acid from profile which had been calculated by PSI-BLAST , *SSconfidence* is the confidences of secondary

structure prediction calculated by PSI-PRED[4]. Next, we simulated folding process with simulated annealing method, started from random coil structure. Folding potential in this process is as follows[5]:

$$V_{total} = V_{vdW} + V_{rama} + V_{HP} + V_{HB} + V_{pairwise}$$

where V_{total} is physically total potential, V_{rama} is ramahcandran potential, V_{tor} is torsions, phi and psi, potential, V_{vdW} is van der Waals interaction, V_{hb} is hydrogen bond interaction, V_{hp} is hydrophobic interaction, and $V_{pairwise}$ is pair-wise interaction.

2. Evaluating TEMPLA models:

Our 3D1D score of 'CIRCLE-FR' was used to evaluate TEMPLA models. If these score were higher than that of our CIECLE server models, we had submitted TEMPLA models ranked by the 'CIRCLE-FR' score which combined with ASA. Otherwise we executed homology modeling method as follows.

3. Constructing structure using FAMS-multi

To obtain the best alignment, 15 homology models which constructed by other server teams of our laboratory, FAMSD, CIRCLE and FUNCTION were collected. These models were scored by the 3D1Dscore of 'CIRCLE10', and then the alignment of the first scored model was used to rebuild model using FAMS-multi (see FAMS-multi abstract).

Results

Now (in 2006/10/03) experimental structures of 80 targets are released. We assessed CA and side chain torsion angles of all server models (TS1).

In the evaluation of CA (GDT_TS) function ranked 12 of 68 servers (all 80 targets). And in the evaluation of $\chi 1$ angle ("correct" side chain residue is within 3.5 in the MaxSub superposition and within 40° from native structure) this FAMS team ranked 5 following ROBETTA, Pmodeller6, FAMSD, and Pcons6. Furthermore in the evaluation of correct side chain number within 2.0 and 1.0 in the MaxSub superposition, this team ranked 4 and 3, respectively.

1. Ogata K. and Umeyama H. (2000) An automatic homology modeling method consisting of database searches and simulated annealing J Mol Graph Model/J Mol Graph Model 18, 258-272, 305-256.
2. Zhou H., Zhou Y. (2004) Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition.Proteins.1;55(4):1005-13.
3. Kim T. Simons, Charlie Strauss and David Baker (2001) Prospects for ab initio Protein Structural Genomics J. Mol. Biol. 306, 1191-1199
4. Jones D.T (1999) Protein secondary structure prediction based on position-specific scoring matrices J. Mol Biol/J. Mol Biol 292, 195-202.
5. Yoshimi Fujitsuka, George Chikenji, Shoji Takada. (2006) SimFold Energy Function For De Novo Protein Structure Prediction: Consensus with Rosetta PROTEINS 62:381-398

FAMS_ACE - 500 models for 100 3D targets

Model selection from server results using original threading(3D1D) program and consensus

Mitsuo Iwadate, Kazuhiko Kanou, Genki Terashi, Daisuke Takaya, Kazuhiro Ohta, Akio Hosoi, Mayuko Takeda-Shitaka and Hideaki Umeyama

Department of Biomolecular Design
School of Pharmaceutical Sciences, Kitasato University
iwadatem@pharm.kitasato-u.ac.jp

"fams_ace" is meta-selector team using all the server models. Concept of the meta-selector that appears with CASP5 2002, has freed many predictors from suffering hardship work. Then we have registered as the manual team, "fams_ace" in CASP7. This team downloaded all the server answers in the CASP7 sight and chose an appropriate model from the submitted model. Again, "fams_ace" has registered in the manual team, but in fact it is a meta-selector or a meta server.

Method

There is 3 points in methodology in choosing the submitted structure.

1.All downloaded server models from CASP7 were rebuilt using homology modeling software, FAMS (3).

2.Many servers that have been registered in the sight of CASP7 had submitted 5 structures. In the process of one structure selection from these 5 structures in each server, the series of the threading software, named "CIRCLE" which was developed in our research group was used. This program separately participates as a server in CASP7. In the process for selection of best one in 5 structure in each team, "fams_ace" used the "CIRCLE".

3.After step 1 mentioned above, best structure are selected using the consensus opinion method. It is clone software of 3D-JULY made by us.

Self-assessment of GDT_TS

Especially many servers refer the third point mentioned above, then the submitted structure of "fams_ace" tends to be similar to structures submitted by many servers.

Now (in 2006/10/03) experimental structures of 80 targets are released. For predicting the target difficulty, we used SVM program using both of PSI-BLAST(1) and SPARKS(2) score and homology percent value. The training data set was CASP6 targets. The accuracy of this prediction was 85% in CASP6 targets. Each target sequence is not divided to domain regions. Total GDT_TS of the 52 targets are 3795.93 by fams_ace. In CASP7 68 severs only Zhang-Server gives the higher point than fams_ace. Accordingly, fams_ace is able to become a top level server.

GDT_TS		
Rank	Score	Server name
1	4908.19	Zhang-Server
2	4815.21	fams_ace
3	4617.83	Pmodeller6
4	4604.31	HHpred2
5	4574.06	CIRCLE
6	4561.34	ROBETTA
7	4539.88	Pcons6
8	4539.05	HHpred3

GDT_TS CM 52 targets			GDT_TS FR 28 targets		
Rank	Score	Server name	Rank	Score	Server name
1	3812.47	Zhang-Server	1	1095.72	**Zhang-Server**
2	3798.93	fams_ace	2	1067.11	Pmodeller6
3	3657.96	CIRCLE	3	1019.28	fams_ace
4	3649.04	UNI-EID_expm	4	977.19	MetaTasser

1. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W. and Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs Nucleic Acids Res/Nucleic Acids Res 25, 3389-3402.
2. Zhou H., Zhou Y. (2004) Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition.Proteins.1;55(4):1005-13.
3. Ogata K. and Umeyama H. (2000) An automatic homology modeling method consisting of database searches and simulated annealing J Mol Graph Model/J Mol Graph Model 18, 258-272, 305-256.

FAMSD - 500 models for 100 3D targets

Homology modeling server providing

side chain models with high accuracy

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

[1] - Department of Biomolecular Design

School of Pharmaceutical Sciences, Kitasato University

kanouk@pharm.kitasato-u.ac.jp

FAMSD is a homology modeling server constructing high accuracy side chain models. In the previous CASP experience (CASP6), FAMSD server used only the alignment score and hadn't used the structure score to select the best model. As a result FAMSD had not selected the good side chain models in the CM/easy category. So we have reconstructed FAMSD server focused on selecting good side chain models for CASP7 using both alignment score and structure score.

Prediction target difficulty

For predicting the target difficulty, we used SVM program. The training data set was CASP6 targets. The accuracy of this prediction was 85% in CASP6 targets. This predicted difficulty was used in the next evaluation step. If difficulty of target sequence is 'CM', the model was constructed according to the following scheme, else the method was same as our CIRCLE server except for no use of the outside server for our research laboratory.

Method description for 'CM' target

1. Selecting alignments: The alignment selection for constructing highly accurate backbone models using homology modeling is as follows. 8 kinds of methods, BLAST [1], PSI-BLAST, PSF-BLAST, RPS-BLAST, IMPALA, FASTA, Pfam and sparks2 [2] were executed for each amino acid sequence of query proteins.

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

For selecting the alignment candidates in 7 kinds of alignment methods (exclude sparks2), the score-function that was constructed by model length, homology% and degree of secondary structure agreement between PSI-PRED and STRIDE was defined:

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

Len is residue length of model protein. Hom indicate homology % value, the ratio between the number of match residues and Len. SS is so called Q3 value, degree of secondary structure agreement between PSI-PRED and STRIDE. k_i are coefficients of each alignment method.

Top 5 alignments ranked by this score and the first scored alignment of sparks2 were selected for homology modeling. Then the each number of selecting alignments (i.e. "top 5" and "first scored") was optimized using CASP6 server models as a training set.

2. Homology Modeling and Refinement models: Models were constructed using selected alignments by homology modeling software FAMS (full automatic modeling system) [3]. After homology modeling, both of Energy Minimization and Molecular Dynamics are applied for refinement models. Especially the hydrogen bonds and collision of model are refined.

3. Selecting good side chain models: All constructed models were evaluated by new 3D1D score 'CIRCLE-10HB'. This score considered hydrogen bonds on the defining 'environment' of each amino acid residues. And this score was optimized for selecting high accurate side chain models using CASP6 server models as a training set.

Results

Now (in 2006/10/03) experimental structures of 80 targets are released. We assessed CA and side chain torsion angles of all server models (TS1).

In the evaluation of CA (GDT_TS) FAMSD ranked 9 of 68 servers (all 80 targets). And in the evaluation of $\chi 1$ angle ("correct" side chain residue is within 3.5 in the MaxSub superposition and within 40° from native structure) this FAMSD team ranked 3 following ROBETTA, Pmodeller6. Furthermore in the evaluation of correct side chain number within 2.0 and 1.0 in the MaxSub superposition, this team ranked 2 and 2, respectively.

1. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W. and Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs Nucleic Acids Res/Nucleic Acids Res 25, 3389-3402.
2. Zhou H., Zhou Y. (2004) Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition.Proteins.1;55(4):1005-13.
3. Ogata K. and Umeyama H. (2000) An automatic homology modeling method consisting of database searches and simulated annealing J Mol Graph Model/J Mol Graph Model 18, 258-272, 305-256.

fams-multi- 509 models for 100 3D/9 TR targets

Homology modeling meta-server using multiple reference proteins

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

[1] - Department of Biomolecular Design
School of Pharmaceutical Sciences, Kitasato University
kanouk@pharm.kitasato-u.ac.jp

Fams-multi is a homology modeling meta-server using all submitted models which were constructed by all the server teams in CASP7. Such submitted models were used to generate the better alignments and rebuilt models by automatic homology modeling software 'FAMS-multi' which is multiple reference-proteins version of FAMS [1]. This server aimed to build models with high quality loop and side chain. In the following, we describe the scheme.

Generating 'best' pairwise alignments

All server models (TS1-TS5) were refined by FAMS for the purpose of removing collision, and these models were evaluated and ranked by the same method as our CIRCLE server. Then top 5 models (in excluding models which hasn't described reference PDB code on the 'PARENT' record) were selected to generate alignments. The alignment was generated by structural alignment between the model and 'parent' PDB using CE program [2].

Constructing models by FAMS-multi

First some reference proteins were chosen based on the certain criteria concerning sequence and structural similarity with 'parent' PDB. Maximum number of the reference proteins is 30. Next, a multiple structural alignment based on the superposition of CA atoms was performed among the reference proteins. For this alignment, the target sequence was put on by sequence alignment generated by CE. This alignment was evaluated to determine if inserted gaps were concentrated in loop and variable regions (VRs), which are

defined by residues having the distance between CA atoms greater than 1.0 Å. Thus, we get a result of multiple alignment between a target and reference proteins.

Using this alignment tertiary structure was constructed with mainly next three steps, CA construction, main-chain construction, side-chain construction. In the each step optimization by the simulated annealing method was executed.

CA construction step: For the initial CA coordinates, first, the weighted average of CA coordinates and the average distance were obtained from pairwase structural alignment based on the superposition of CA atoms between the target and reference proteins. Next, simulated annealing optimized the coordinates of CA atoms.

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

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

Evaluating models

5 models constructed using FAMS-multi were selected in the same method as our CIRCLE server.

Refinement experiment

Fams-multi had participated in refinement experiment using Energy minimize & Molecular dynamics. Refined models are correctly revised for hydrogen bonds, main-chain torsion angles, side-chain torsion angles and the decreasing collision between hydrophobic atoms.

Results

The model 2 of fams-multi on T0288 was adopted as the initial strucutre of refinement experiment. This model scored second according to the GDT_TS score. Comparing with all the server teams (TS1) in 2006/10/03, in the GDT_TS estimation and $\chi 1$ estimation this meta-server fams-multi team ranked 2 and 2, respectively. In the case of CM (see FAMSD abstract) this team ranked 2 and 1, respectively.

1. Ogata K. and Umeyama H. (2000) An automatic homology modeling method consisting of database searches and simulated annealing J.Mol Graph Model/J Mol Graph Model 18, 258-272, 305-256.
2. Shindyalov I.N., Bourne P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Engineering 11(9) 739-747.

Feig - 469 models for 99 3D targets

Sampling and Scoring Strategies in an Iterated Protocol for Protein Structure Prediction

Katarzyna Maksimiak[1]*, Andrew Stumpff-Kane[1] and Michael Feig[1,2]

[1]*Department of Biochemistry and Molecular Biology*

[2]*Dept. of Chemistry, Michigan State University, East Lansing, MI 48824; USA*

feig@msu.edu

We have developed an iterated protocol for protein structure prediction. In this protocol, we seek to build homology models by, first, generating a diverse set of potential alignments for each target; creating models from each alignment using loop modeling to fill gaps as necessary; and evaluating the models using various scoring functions combined with statistical techniques to reduce the effect of noise. In appropriate cases we seek to refine the models further, employing iterative rounds of lattice modeling or, in cases of high homology, normal mode-based sampling to generate additional sample conformations. In particular, initial templates were obtained using both sequence- and fold-recognition methods; then, for each template, an ensemble of "suboptimal" alignments was generated using PROBA. To score the models we used a combination of the knowledge-based scoring functions DFIRE[1], Verify3D[2], RAPDF[3] and ProsaII[4]; together with clustering and a correlation-based method for reducing noise[5]. In the refinement stage, we employed MONSSTER[6] to generate samples for medium-homology templates and normal-mode-based sampling, in connection with DFIRE, for high-homology templates.

1. Zhang C., Liu S., Zhu Q.Q. & Zhou Y.Q. (2005) A knowledge-based energy function fro protein-ligand, protein-protein, and protein-DNA complexes, J. Med. Chem. 48, 2325-2335.
2. Luthy R., Bowie J.U. & Eisenberg D. (1992) Assessment of protein models with three-dimensional profiles, Nature 356 83–85.
3. Samudrala R. & Moult J. (1998) An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction, J. Mol. Biol 275, 895-916.

4. Sippl M.J. (1993) Recognition of errors in three-dimensional structures of proteins, Proteins 17, 355-362.
5. Stumpff-Kane A. & Feig M. (2006) A correlation-based method for the enhancement of scoring functions on funnel-shaped energy landscapes, Proteins 63, 155-164.
6. Skolnick J., Kolinski A. & Ortiz A.R. (1997) MONSSTER: A method for folding globular proteins with a small number of distance restraints, J. Mol. Biol 265**,** 217-41.

fleil - 311 models for 63 3D targets

Comparative modeling with all-atom refinement using molecular dynamics simulation

S. Fuchigami[1], T. Amemiya[1], S. Oomori[1] & R. Koike[1,2]
[1] - International Graduate School of Arts and Sciences, Yokohama City University, [2] - Global Scientific Information and Computing Center, Tokyo Institute of Technology
sotaro@tsurumi.yokohama-cu.ac.jp

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

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

As pointed out by Misura et al.[4], models produced with MODELLER generally contain atomic clashes, which are detected by using the program Probe[5]. To remove the atomic clashes in the models, we carried out energy minimization by steepest descents using the MD program system, MARBLE[6], with the CHARMM22 force field for proteins[7] and the CMAP correction for peptide backbone ϕ, ψ dihedral crossterms[8]. Consequently the clashes were considerably reduced to the same extent or less than observed in native crystal structures.

In order to sample possible conformations of the target proteins at atomistic level, we performed MD simulation in NPT ensemble with explicit water, started with the energy-minimized structures, using the MARBLE[6] with the same CHARMM force filed parameters as mentioned above. The initial structures were dissolved in water molecules with the addition of counter ions to neutralize the net charges of the system. The temperature and pressure of the system were set at 300 K and 1 atom, respectively. Water molecules and hydrogen-containing group (e.g. CH_3, NH_2, OH, etc.) were treated as rigid bodies (partial rigid-body method), enabling to use a 2.0 fs time step. Coulombic interactions were evaluated using the particle-mesh Ewald method[9]. For some targets, additional refinements were carried out using simulated annealing to relax the sampled conformations of the target, especially fluctuating loops

Submitted models were chosen from a set of models generated using different templates and alignments based on complete-linkage clustering, ranking of radii of gyration, or visual inspection.

1. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W. & Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Acids Res. 25, 3389-3402.
2. Jones D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292, 195-202.
3. Šali A. & Blundell T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234, 779-815.
4. Misura K.M.S., Chivian D., Rohl C.A., Kim D.E. & Baker D. (2006) Physically realistic homology models built with ROSETTA can be more accurate than their templates. Proc. Natl Acad. Sci. 103, 5361-5366.
5. Word J.M., Lovell S.C., LaBean T.H., Taylor H.C., Zalis M.E., Presley B.K., Richardson J.S. & Richardson D.C. (1999) Visualizing and Quantifying Molecular Goodness-of-Fit: Small-probe Contact Dots with Explicit Hydrogen Atoms. J. Mol. Biol. 285, 1709-1731.
6. Ikeguchi M. (2004) Partial Rigid-Body Dynamics in NPT, NPAT and NPT Ensembles for Proteins and Membranes. J. Comput. Chem. 25, 529-541.
7. MacKerell A.D., Jr., Brooks B., Brooks C.L., III, Nilsson L., Roux B., Won Y. & Karplus M. (1998) CHARMM: The Energy Function and Its Parameterization with an Overview of the Program. in The Encyclopedia of Computational Chemistry edited by Schleyer,P.v.R. et al., (John Wiley & Sons, Chichester, 1998), Vol. 1, pp. 271-277.
8. MacKerell A.D., Jr. (2004) Empirical Force Fields for Biological Macromolecules: Overview and Issues. J. Comput. Chem. 25, 1584-1606.
9. Essmann U., Perera L., Berkowitz M.L., Darden T., Lee H. & Pedersen L.G. (1995) A smooth particle mesh Ewald method. J. Chem. Phys. 103, 8577-8593.

Floudas - 150 models for 30 3D/1 TR targets

First Principles Protein Structure Prediction

S.R McAllister[1], R. Rajgaria[1] and C.A. Floudas[1]
[1] - *Princeton University*
floudas@titan.princeton.edu

We present the development and application of ASTRO-FOLD, a novel and complete approach for the first principles prediction of protein structures given only the amino acid sequences of the proteins[1]. The approach exhibits many novel components and the merits of its application have been examined for a suite of protein systems, including targets from several CASP experiments. The main thrusts of this approach are α-helical prediction through detailed energy calculations, a global optimization formulation for the β-sheet prediction, the derivation of secondary structure restraints and loop modeling, and the application of a hybrid global optimization algorithm to tertiary structure prediction.

The first stage involves the identification of helical segments and is accomplished by first partitioning the amino acid sequence into overlapping oligopeptides[2]. These oligopeptides are modeled at the atomistic level using the ECEPP/3 force field, where an ensemble of low energy conformations is generated. Given this ensemble, the free energies are calculated, including entropic, cavity formation, polarization and ionization contributions for each oligopeptide. The helical propensity for each residue is then identified using equilibrium occupational probabilities of helical clusters. However, due to the time constraints of predictions, information from the PSI-PRED server has been used as the base prediction, supplemented by this first stage approach.

The second stage focuses on the prediction of β-sheet and disulfide bridge topology by first postulating a β-strand superstructure that encompasses all alternative beta-strand arrangements[3]. This β-strand superstructure is used to model the hydrophobic driving force important for β-structure formation through an integer linear optimization model originally developed in the area of process synthesis of chemical systems. The resulting optimization model is solved to maximize the hydrophobic contact energy, thus providing a rank ordered list of preferred hydrophobic residue contacts, beta strand topologies and disulfide bridge connectivities. Due to time constraints, only a few predicted β-sheet topologies were selected for further study.

The third stage involves the derivation of restraints based on helical and beta-sheet predictions in the form of dihedral angle and atomic distance restraints to enforce the predicted secondary and tertiary arrangements. For entirely α-helical proteins, a novel optimization framework has been used to predict topological contacts and generate interhelical distance restraints between hydrophobic residues[4]. Additional restraints are determined for the intervening loop residues connecting helical and strand regions through dihedral angle sampling and a novel clustering approach[5].

The fourth and final stage of the approach involves the prediction of the tertiary structure of the full protein sequence. The problem formulation, which relies on dihedral angle and atomic distance restraints introduced from the previous stages as well as detailed atomistic energy modeling, represents a nonconvex constrained global optimization problem, which is solved through the combination of a deterministically based global optimization approach, the αBB, and torsion angle dynamics. The use of the αBB global optimization algorithm guarantees convergence to the global minimum solution by a convergence of upper and lower bounds on the potential energy minimum. By applying torsion angle dynamics (TAD) as an initialization step and a stochastic global optimization method such as conformation space annealing (CSA), the difficulty of converging to the global minimum is significantly reduced by the quick determination of low energy conformations[6]. This hybrid approach was run only for a limited time due to the deadlines imposed.

1. Klepeis J.L. & Floudas C.A. (2003) A combinatorial and global optimization framework for ab initio prediction of three-dimensional structures of proteins from the amino acid sequence. Biophys. J. 85, 2119-2146.
2. Klepeis J.L. & Floudas C.A. (2002) Ab initio prediction of helical segments in polypeptides. J. Comput. Chem. 23, 245-266.
3. Klepeis J.L. & Floudas C.A. (2003) Prediction of beta-sheet topology and disulfide bridges in polypeptides. J. Comput. Chem. 24, 191-208.
4. McAllister S.R., Mickus B.E., Klepeis J.L. & Floudas C.A. (2006) A novel approach for alpha-helical topology prediction in globular proteins: generation of interhelical restraints. Prot. Struct. Funct. Bioinf., accepted for publication.
5. Mönnigmann M. & Floudas C.A. (2005) Protein loop structure prediction with flexible stem geometries. Prot. Struct. Funct. Bioinf. 61, 748-762.
6. Klepeis J.L, Pieja M.T. & Floudas C.A. (2003) Hybrid global optimization algorithms for protein structure prediction: alternating hybrids. Biophys. J. 84, 869-882.

FOLDpro (server, domain) - 600 models for 100 3D/ 100 DP targets

Domain Prediction Using FOLDpro and DOMpro

Jianlin Cheng, Mike Sweredoski, and Pierre Baldi
Institute for Genomics and Bioinformatics, School of Information and Computer Science
University of California Irvine, CA 92697

In CASP7, we combine fold recognition approach [1] and *ab initio* approach [3] together to predict protein domains. For a query protein, our domain prediction server (FOLDpro) first ranks templates by using support vector machine (SVM) to integrate alignment and structural features of query-template pairs [1]. If the top template is significant enough (SVM score > -0.5), FOLDpro generates a 3D model for the query protein from the template and uses PDP [2] to parse the model into domains. If domains generated by PDP do not cover the whole query sequence, FOLDpro uses a post processing step to assign uncovered regions to adjacent domains. If no significant template is found, FOLDpro invokes DOMpro [3], an *ab initio* domain predictor using neural networks, profiles, and structural features, to predict domains.

1. Cheng J., and Baldi P. (2006) A Machine Learning Information Retrieval Approach to Protein Fold Recognition. Bioinformatics, vol. 22, no. 12, pp. 1456-1463.
2. Alexandrov N. and Shindyalov I. (2003) PDP: protein domain parser. Bioinformatics, vol. 19, pp. 429-430.
3. Cheng J., Sweredoski M., and Baldi P. (2006) DOMpro: Protein Domain Prediction Using Profiles, Secondary Structure, Relative Solvent Accessibility, and Recursive Neural Networks. Data Mining and Knowledge Discovery, vol. 13, no. 1, pp. 1-10.

FORTE1 - 499 models for 100 3D targets

FORTE1: A Profile-Profile Comparison Method for Fold Recognition

K. Tomii
Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, 2-42 Aomi, Koto-ku, Tokyo, Japan
k-tomii@aist.go.jp

We have built automated fold recognition server, FORTE1, based on a profile-profile comparison method since the CASP5 experiment. FORTE is an abbreviation for "FOld Recognition TEchnique". The server[1] is publicly available for academic use. This approach has also been applied to protein structure prediction of the CASP7 targets.

The FORTE1 system uses position-specific score matrices (PSSMs) of both the query and templates as profiles. It identifies proper templates and produces profile-profile alignments of a target and templates. To calculate PSSMs of both the query and templates, PSI-BLAST[2] iterations are performed a maximum of 20 times with the NCBI non-redundant database. The amino acid sequences of templates are derived from the ASTRAL[3] 40% identity list and selected PDB[4] entries that are not registered in the SCOP[5] database. Furthermore, the full-length sequences, which are divided into structural domains in SCOP, are also prepared.

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

1. Tomii K., & Akiyama Y. (2004) FORTE: a profile-profile comparison tool for protein fold recognition. Bioinformatics 20, 594-595.
2. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., & Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389-3402.

3. Chandonia J.M., Hon G., Walker N.S., Lo Conte L., Koehl P., Levitt M. & Brenner S.E. (2004) The ASTRAL compendium in 2004. Nucleic Acids Res. 32, D189-D192.
4. Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N. & Bourne P.E. (2000) The Protein Data Bank. Nucleic Acids Res. 28, 235-242.
5. Andreeva A., Howorth D., Brenner S.E., Hubbard T.J.P., Chothia C. & Murzin A.G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res. 32, D226-D229.

FORTE2 - 499 models for 100 3D targets

FORTE2: Automated Fold Recognition Server with Enhanced Profile Library

K. Tomii

Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, 2-42 Aomi, Koto-ku, Tokyo, Japan
k-tomii@aist.go.jp

To elucidate effects including very distantly related sequences into profiles for alignment accuracy, as well as sensitivity and selectivity of fold recognition, we have constructed our new server: FORTE2 (FORTE is an abbreviation for "FOld Recognition TEchnique"). Its system uses the same protocol as FORTE1[1]. It has enriched profiles by incorporating highly diverged sequences detected by FORTE1 into the sets of sequences that are gathered by PSI-BLAST[2]. We have found that FORTE2 can detect relationships between proteins that are different from those detected by FORTE1 through the CASP6 experiments[3].

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

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

1. Tomii K., & Akiyama Y. (2004) FORTE: a profile-profile comparison tool for protein fold recognition. Bioinformatics 20, 594-595.
2. Tomii K., Hirokawa T. & Motono C. (2005) Protein structure prediction using various profile libraries and 3D verification. Proteins 61, 114-121.
3. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W. & Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389-3402.
4. Murzin A., Brenner S.E., Hubbard T. & Chothia C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536-540.

FPSOLVER_SERVER- 436 models for 96 3D targets

Ab-initio Protein Structure Prediction Using Backtrack Search

Miguel Bugalho[1] and Arlindo L. Oliveira[1]

[1] *INESC-ID/IST Lisboa Portugal*
mmfb@algos.inesc-id.pr

In this approach we used a search based method for the *ab-initio* prediction of tertiary protein structure. We worked with a fixed set of *Phi/Psi* dihedral angles for each amino acid in the protein chain. The exact set of values used for each dihedral angle depends of the type of amino acid.

The amber99[1] Van der Walls energy is used to detect and avoid clashes in the structure and the radius of gyration is used to bound the search, since one is aiming for a compact structure. The side chains are set using a rotamer library[2].

Our search method is guided by a statistical energy function generated from the proteins in the Whatif[3] database (2004). Instead of using the function for evaluating each pair of dihedral angles, we generate several fragments of N amino acids and then use the function to choose one of those fragments (the fragment technique is successfully used by many other methods, like, for example, the ROSETTA[4] method). This technique avoids the problem of not having enough information to make a decision, as it would have happened if we position one amino acid at a time.

The proposed approach is not guaranteed to find a structure equal or arbitrarily near the target protein, but a nearby solution (lower than 3 angstroms of

RMSD) is achievable in many cases. Since the solution can than be refined to a nearer solution by other techniques, the problem is then to find a good search strategy in the exponentially large search space.

Some search approaches have been described in the literature, normally to sample the conformational space for small proteins or parts of proteins. These approaches commonly use methods like filtering, perform sorting using scoring functions, or use clustering for choosing some of the conformations in the sample. The proposed approach tries to use these techniques to guide the search and create near native conformations, instead of just conformational samples.

Some of the most successful algorithms for *ab-initio* prediction use simulated annealing for searching chain conformations. This simple method has obtained good results in many minimization problems.

Although the search based approach is *per se* also a simple method, the search method used, the heuristics, the pruning strategies and all other techniques that can be used in conjunction with the search provide much more room for adapting the method to the protein structure prediction problem. This type of adaptation is normally made in the simulated annealing method through the modification of the function to minimize. Unfortunately, the problem of finding a good heuristic related with the energy that can be use to effectively guide the search is still far from being solved.

The proposed approach uses a technique that is not so heavily dependent on the *fitness* function, and can therefore accommodate other information. Preliminary results show that this approach is able to generate good solutions for very small proteins, but that the search techniques still need improvements to make the method applicable to larger proteins.

1. Cornell D., Cieplak P., Bayly I., Gould I.R., Merz K.M., Ferguson D.M., Spellmeyer D.C., Fox T., Caldwell J.W. and Kollman P.A. (1995) A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids and Organic Molecules. J. of the American Chemical Society 117, 5179-5197.
2. Dunbrack R.L., Jr. and Karplus M. (1993) Backbone-dependent Rotamer Library for Proteins: Application to Side-chain prediction. J. Mol. Biol. 230, 543-574.
3. Vriend G. (1990) WHAT IF: A molecular modeling and drug design program., J. Mol. Graph. 8, 52-56.
4. Simons K.T., Kooperberg C., Huang E., Baker D. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulate anealing and Bayesian scoring functions. J Mol Biol 268, 209-25.

FUNCTION - 500 models for 100 3D targets

Building many models using FAMS and selecting model with Special scoring Function

Mitsuo Iwadate, Kazuhiko Kanou, Genki Terashi, Daisuke Takaya, Kazuhiro Ohta, Akio Hosoi, Mayuko Takeda-Shitaka and Hideaki Umeyama

Department of Biomolecular Design
School of Pharmaceutical Sciences, Kitasato University
iwadatem@pharm.kitasato-u.ac.jp

FUNCTION is a homology-modeling server constructing high accuracy side chain models. In the previous CASP experience (CASP5 in 2002), FAMSD server used both of the alignment score and the structure score to select the best model. As a result FAMSD in CASP5 had selected the good side chain models in the CM/easy category in CAFASP3 section. So we have reconstructed FAMSD server focused on selecting good side chain models for CASP7 using both alignment score and the same structure score. Additionally in CASP6 SPARKS2[2] calculates good structure, and then the software was also used.

Prediction target difficulty

For predicting the target difficulty, we used SVM program (http://www.smartlab.dibe.unige.it/). The training data set was CASP6 targets. The accuracy of this prediction was 85% in CASP6 targets. This predicted difficulty was used in the next process.

Method

Alignments:

The alignment selection for constructing highly accurate backbone models using homology modeling is as follows. 8 kinds of methods, BLAST [1], PSI-BLAST, PSF-BLAST, RPS-BLAST, IMPALA, FASTA, Pfam and sparks2 [2] were executed for each amino acid sequence of query proteins.

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

Modeling:

For all the alignments of E-value<=0.1 were built structural models.

Top 5 alignments ranked by this scoring function and the first scored alignment of sparks2 were selected for homology modeling.

Model selection

For selecting the model candidates in 7 kinds of alignment methods (in excluding sparks2), the score-function that was constructed by model length, e value and degree of secondary structure agreement between PSI-PRED and STRIDE was defined:

$$score = f(e, Len, SS, enesosui)$$

Len is residue length of model protein. *e* indicate e value of BLAST or FASTA, the ratio between the number of match residues and *Len*. *SS* is so called Q3 value, degree of secondary structure agreement between PSI-PRED and STRIDE. *enesosui* is degree of hydrophobic interaction.

Results

Now (in 2006/10/03) experimental structures of 80 targets are released. We assessed CA and side chain torsion angles of all server models (TS1).

In the evaluation of CA (GDT_TS) FUNCTION ranked 21 of 68 servers (all 80 targets). And in the evaluation of χ1 angle ("correct" side chain residue is within 3.5 in the MaxSub superposition and within 40° from native structure) this FUNCTION team ranked 8 following ROBETTA, Pmodeller6, FAMSD, Pcons6, FAMS, Zhang-Server, CIRCLE. Furthermore in the evaluation of correct side chain number within 2.0 in the MaxSub superposition, this team ranked 6 following ROBETTA, FAMSD, Pmodeller6, FAMS, Pcons6.

1. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W. and Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs Nucleic Acids Res/Nucleic Acids Res 25, 3389-3402.
2. Zhou H., Zhou Y. (2004) Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition.Proteins.1;55(4):1005-13.
3. Ogata K. and Umeyama H. (2000) An automatic homology modeling method consisting of database searches and simulated annealing J Mol Graph Model/J Mol Graph Model 18, 258-272, 305-256.

GeneSilico - 616 models for 96 3D/ 92 DP / 91 QA targets

Identification and refinement of potential errors in protein structures.

M. Boniecki[1], M. Pawlowski[1,2], M.J. Gajda[1,3], M.J. Pietal[1,2], A. Kaminski[1,2], A. Obarska[1,2], M. Fijalkowski[1], M. Feder[1], G. Papaj[1,2], Tkaczuk K.L.[1,4], Lopez Munoz L.[5], J. Orlowski[1], M.A. Kurowski[1], J.M. Sasin[1], J.M. Bujnicki[1,*], and J. Kosinski[1]

[1] *International Institute of Molecular and Cell Biology, Trojdena 4, 02-109 Warsaw, Poland*

[2] *Institute of Biochemistry and Biophysics PAS, Pawinskiego 5A, 02-106 Warsaw, Poland*

[3] *Faculty of Mathematics and Information Science, Warsaw University of Technology Warsaw, Poland*

[4] *Institute of Technical Biochemistry, Technical University of Lodz, Stefanowskiego 4/10, 90-924 Lodz, Poland*

[5] *ICM, Warsaw University, Pawinskiego 5A, 02-106 Warsaw, Poland. * iamb@genesilico.pl*

In CASP7, we used the strategy developed in the course of CASP5 and CASP6, with several important modifications and the shift of emphasis from building of models to model quality assessment and refinement.

First, we upgraded some of our old tools. We developed a new version of the GeneSilico meta-server [1], which now contains multiple methods for protein domain identification, and residue-level prediction of disorder, secondary structure and solvent accessibility. We have also developed a set of new tools for visualization and clustering of protein contact maps (e.g. PROTMAP2D, M.J.P., Irina Tuszynska and J.M.B., manuscript in preparation), which have been used to identify independently folded protein domains. We incorporated new fold-recognition algorithms into the meta-server, in particular the HHSearch method [2] for profile-profile alignments and a number of tools for post-processing of crude 3D models, including the creation of hybrid models according to the 'Frankenstein's monster' philosophy [3; 4; 5]

Second, we developed a new 'meta-server' for model quality assessment (Meta-MQAP), which uses several primary MQAPs to derive a score that represents a predicted deviation (in Angstroms) of individual residues in the model with respect to their counterparts in the (unknown) native structure. According to our benchmarks, Meta-MQAP is significantly better from all primary methods (M.P., Ryszard Matlak, and J.M.B., in preparation).

Third, we developed a system for data management called UniMod, which we use as a framework with a common WWW interface to run different in-house and third-party methods using common formats (an example is the possibility

to run MODELLER with a project file generated in SwissPDBViewer, with additional restraints e.g. on predicted secondary structure), to store the results in a database and to pass the files between different programs. In particular, UniMod has been used to generate models and to score them according to MetaMQAP.

The evaluated models obtained from the meta-server and the UniMod pipeline were used as a source of spatial restraints for simulations, in a similar manner to the strategy used by the Kolinski-Bujnicki group in CASP6, but with a few important differences. In particular, for de-novo folding based on restraints we used a real-space method REFINER [6] with a new potential of mean force. In confident models (those based on highly similar templates or with very high average MetaMQAP scores), we refined only the regions with very poor scores or high diversity between different model variants. The REFINER and MetaMQAP scores were used as the primary criteria for selection of the final models. For difficult targets we generated additional de-novo models using REFINER [6] and ROSETTA [7] and clustered them together with the fold-recognition models to identify the most frequently occurring conformations with low energy. Members of the selected clusters were then used as a source of spatial restraints (derived from residues with good MetaMQAP scores) to generate a representative structure with REFINER. Because REFINER uses a reduced representation and the reconstructed full-atom models sometimes exhibit minor stereochemical errors, the final models were 'idealized' with MODELLER [8].

1. Kurowski M.A. and Bujnicki J.M. (2003) GeneSilico protein structure prediction meta-server, Nucleic Acids Res., 31, 3305 - 3307.
2. Söding J. (2005) Protein homology detection by HMM–HMM comparison Bioinformatics., 21, 951 - 960.
3. Kosinski J., Cymerman I.A., Feder M., Kurowski M.A., Sasin J.M., Bujnicki J.M. (2003) A ldquoFRankenstein's monsterrdquo approach to comparative modeling:Merging the finest fragments of Fold-Recognition models and iterativemodel refinement aided by 3D structure evaluation, Proteins: Structure, Function, and Genetics., 53, 369-379.
4. Kosinski J., Gajda M.J., Cymerman I.A., Kurowski M.A., Pawlowski M., Boniecki M., Obarska A., Papaj G., Sroczynska-Obuchowicz P., Tkaczuk K.L., Sniezynska P., Sasin J.M, Augustyn A., Bujnicki J.M., Feder M. (2005) FRankenstein becomes a cyborg: The automatic recombination and realignment of fold recognition models in CASP6 Proteins: Structure, Function, and Bioinformatics, 61, S7, 106-113.
5. Kolinski A., Bujnicki J.M. (2005) Generalized protein structure prediction based on combination of fold-recognition with de novo folding and evaluation of models. Proteins, 61, S7, 84-90.
6. Boniecki M., Rotkiewicz P., Skolnick J., Kolinski A. (2003) Protein fragment reconstruction using various modeling techniques. J Comput Aided Mol Des., 725-38.
7. Simons K.T., Kooperberg C., Huang E., Baker D. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol., 268(1), 209-25.
8. Sali A., Blundell T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol., 234(3), 779-815.

Gerloff

A simplified representation of electrostatic model surfaces for addressing protein-protein interaction problems

D.L. Gerloff[1], S. Ali[2] and R. König[3]

[1] – UC Santa Cruz, [2] – Univ. of Edinburgh, [3] – Univ. of Applied Sciences Bingen

d.gerloff@ed.ac.uk

The electrostatic surface properties of protein structures can provide clues towards the interactions with other molecules in which they engage. Particularly the comparison of the electrostatic surfaces of several homologous proteins can prove interesting in this regard – and their structures can often be modelled using comparative modelling techniques. Even if one takes into consideration that the surface properties of a comparative model can merely be an approximation of those of the true structure, it should be of great interest in the context of various protein-protein interaction problems to be able to compare electrostatic model surfaces systematically. Only very few methods exist currently for undertaking such comparisons.

We have developed a novel way of simplifying the comparison of the electrostatic molecular surfaces of proteins to comparing **1-D "electrostatic surface profiles"**. In these surface profiles the electrostatic surface charge of each protein is essentially apportioned to its individual residues. On our poster we will show two examples of how comparisons of the electrostatic model surfaces of homologous proteins using surface profiles may prove useful in protein-protein interaction questions (see below). While the simplified 1-D representation proposed here will necessarily mean a "neglect" of more fine-grained information the profile format offers many advantages over the classic 3-D format of electrostatic potential surfaces. Most obviously analyses as those described below could be combined more easily with multiple sequence analyses of various kinds, e.g. correlated mutation analysis between potential partner proteins in protein-protein interactions.

Binding site prediction - Complement Receptor 1 (CR1): Our results indicate that systematic comparisons of surface profiles are helpful for pinpointing functionally important domains within a set of homologous domains in the

same protein (e.g. in the human immune-regulating protein Complement Receptor 1). While CR1 is known to be involved in protein-protein interactions with several partners at several sites along its length (1998 aa), not all partners are known and the location of the binding sites on different domains, with respect to their common structural scaffold, can differ. Comparing the surface profiles of the models of the 30 homologous domains in CR1[1] to each other, by reference to their sequence similarity, suggests which domain surfaces seem to have changed more than would be expected – which may reflect the acquisition of new interaction partners during evolution. While some specifics of how best to select the most "outstanding" domains remain to be worked out better before this screening approach can be generalised fully, our results agree well with visual inspection of GRASP[2] pictures and can be compared to those of other methods for electrostatic surface comparison. While experimental information about interactions between CR1 and other proteins is scarce, the domains pinpointed by our comparisons seem to be involved in such interactions and our results are compatible with the current biological knowledge.

Partner prediction - CDK-cyclin homologues in *Arabidopsis thaliana*: Where families of paralogous proteins exist (within the same species), not every member of the one protein family will necessarily interact with every member of the other protein family. In a previous project[3], we investigated the potential of a molecular docking approach with modelled protein structures for answering the question which are the most plausibly interacting partners in CDK-cyclin like transient complexes between the approximately 35 CDK and 50 cyclin homologues in *Arabidopsis thaliana* (*At*). In contrast to the interaction problem described above, the three-dimensional orientation of the two partner proteins in putative complexes can be assumed to be similar in all complexes formed. Intersecting the results of molecular docking using ZDOCK[4] with electrostatic complementarity analysis using the program MOLSURFER[5] suggested 19 most likely interacting CDK-cyclin pairs out of the 1188 possible pairs. An alternative prediction method for this problem is being derived in which all possible CDK-cyclin combinations were modelled and the electrostatic surface profiles of their subunits examined for complementarity over the range of their interacting residues. While there are hardly any wet-lab data against which to validate the results by both approaches, their predictions can be compared to one another.

1. Soares D.C., Gerloff D.L., Syme N.R., Coulson A.F.W., Parkinson J. & Barlow P.N. (2005) Large-scale modelling as a route to multiple surface comparisons of the CCP module family. Prot. Eng. Des. Sel. 18, 379-88.
2. Nicholls A., Sharp K.A., & Honig B. (1991) Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. Proteins 11, 281-296.
3. Quan X., Doerner P. & Gerloff D.L. (2006) in preparation.
4. Chen R., Li L. & Weng Z. (2003) ZDOCK: an initial-stage protein-docking algorithm. Proteins 52, 80-87.
5. Gabdoulline R.R., Wade R.C. & D. Walther (2003) MolSurfer: a macromolecular interface navigator. Nucl. Acids Res. 31, 3349-3351.

gtg - 201 models for 45 3D targets

Transitive alignment of distantly-related proteins

Christopher Wilton[1], Swapan Mallick[1], Liisa Holm[1]

[1]-Institute of Biotechnology, University of Helsinki

liisa.holm@helsinki.fi

The gtg-server used during the CASP7 prediction season was buggy and therefore the predictions submitted do not represent the real performance of the method.

The method was developed on a different computer from the server computer. A wrong version of a database index file was transferred to the server. As a consequence of this, only a small part of the database was visible to the search engine.

We noticed the bug only after the CASP7 prediction season. A set of predictions generated using a correctly functioning gtg-server is available from http://ekhidna.biocenter.helsinki.fi/casp7/.

The Global Trace Graph (GTG) v2 was used to find PDB templates. The search uses transitive alignment of distantly-related proteins using a weighted Consistency-Motif score. The concepts of consistency and transitive alignment are described in detail in [1].

The method generates a local sequence alignment (AL format) between the prediction target and a protein in the PDB. Alignment scores above 1000 generally indicate a homologous relationship. Very distant homologues are recognized with scores between 200 and 1000, with reliability around 50 % towards the lower end of scores. Since the method uses transitivity to find the relationship between the query sequence and all proteins that have a structure associated with them, a score of less than 200 is a prediction of a new fold.

This is a purely sequence based method which uses no structural information to generate alignments.

1. Heger A., Lappe M. & Holm L. (2003) Accurate detection of very sparse sequence motifs. RECOMB 2003: Proceedings of the 7th Annual International Conference on Research in Computational Biology, Eds Miller et al. Association for Computing Machinery, New York, NY. pp. 139-147.

HHpred1 – 298 models for 100 3D/98 DP targets

Homology-based structure, function, and domain prediction by HMM-HMM comparison

Johannes Soding[1]
[1]*-Max-Planck-Institute for Developmental Biology*
johannes.soeding@tuebingen.mpg.de

HHpred1 is an automatic version of our structure and function prediction server HHpred (http://hhpred.tuebingen.mpg.de/) and is the simplest of four related servers participating in CASP7 (HHpred1 to 3, BayesHH). It uses HMM-HMM comparison with integrated secondary structure comparison, correlation scoring, and a novel local HMM-HMM maximum a-posteriori probability (MAP) alignment scheme. Its main difference from a HHpred.2 in CASP6 is its use of a weekly updated HMM database derived from the PDB instead of SCOP and the use of a local MAP alignment scheme.

The tertiary structure prediction proceeds in four steps:

1. HHpred builds a multiple alignment from the target sequence with PSI-BLAST (1) (up to 8 rounds with E-value threshold 1E-3). PSIPRED (2) is used for secondary structure prediction.

2. The alignment is converted to an HMM and compared with a database of HMMs derived from representative sequences in the PDB (70% maximum sequence identity) using the HHsearch software (3) in local Viterbi alignment mode.

3. The top alignments are redetermined using the local MAP scheme.

4. The alignment for the best Viterbi match is submitted to MODELLER (3) to generate a homology model.

For function prediction, the target HMM is compared with the PDB and the Interpro database (5) using HHM-HMM comparison. Mappings to GO numbers are either provided by the GOA (6) and InterPro databases or, if these are not available, assigned by weighted word counts In this case, each word in the PFAM or PDB name and description text casts votes for GO terms containing this word. Words are wwighted depending on their frequency in the GO definition file and on a word frequency table for standard english.

For homology-based domain prediction, the target HMM is compared with the SCOP (7) and Pfam (8) databases using HHsearch in local Viterbi mode. The top alignments are realigned in global Viterbi mode and the aligned regions of the top-scoring hits overlapping not more than 20 residues and possessing at least 50 aligned residues define the domain boundaries.

1. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389-3402.
2. Jones D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol. 292:195-202.
3. Söding J. (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics. 21:951-960.
4. Sali A., Blundell T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol. 1993 234:779-815.
5. http://www.ebi.ac.uk/interpro/
6. http://www.ebi.ac.uk/GOA/
7. http://scop.mrc-lmb.cam.ac.uk/scop/
8. http://www.ebi.ac.uk/interpro/

HHpred2 - 100 models for 100 3D targets

Homology-based structure prediction by HMM-HMM comparison and multiple template selection

Johannes Soding[1]
[1]*-Max-Planck-Institute for Developmental Biology*
johannes.soeding@tuebingen.mpg.de

HHpred2 is an automatic version of our structure and function prediction server HHpred (http://hhpred.tuebingen.mpg.de/) and is one of four related servers participating in CASP7 (HHpred1 to 3, BayesHH). It uses HMM-HMM comparison with integrated secondary structure comparison, correlation scoring, a novel local HMM-HMM maximum a-posteriori probability (MAP) alignment scheme, and multiple template selection. It typically returns a 3D model within ~15 minutes.

The tertiary structure prediction proceeds in five steps (Steps 1, 2, and 4 are the same for HHpred1):

1. Build a multiple alignment from the target sequence with PSI-BLAST (1) (up to 8 rounds with E-value threshold 1E-3). PSIPRED (2) is used for secondary structure prediction.

2. The alignment is converted to an HMM and compared with a database of HMMs derived from representative sequences in the PDB, using the HHsearch software (3) in local Viterbi alignment mode.

3. The top 20 matches are clustered by UPGMA into a forest of separate trees, based on the structure comparison scores of TM-align (4). The clustering stops when the highest average pairwise TM-score drops below 0.7. For each tree, a

multiple structural alignment is calculated with MUSTANG (5). The corresponding PSI-BLAST alignments are merged into a super-alignment in a master-slave fashion and an HMM is generated. The target HMM is compared with these HMMs and the best match defines a set of templates.

4. The top-scoring alignment with these templates is redetermined using the local MAP scheme.

5. MODELLER (6) is used to generate a homology model from this multiple template alignment.

1. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389-3402.
2. Jones D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol. 292:195-202.
3. Söding J. (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics. 21:951-960.
4. Zhang Y., Skolnick J. (2004) Scoring function for automated assessment of protein structure template quality. Proteins. 57:702-710.
5. Konagurthu A.S., Whisstock J.C., Stuckey P.J., Lesk A.M. (2006) MUSTANG: a multiple structural alignment algorithm.Proteins. 64:559-574.
6. Sali A., Blundell T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol. 1993 234:779-815.

HHpred3 - 300 models for 100 3D/100 DP/100 FN targets

Homology-based structure, function, and domain prediction by HMM-HMM comparison, multiple template selection, and intermediate profile search

Johannes Soding[1]

[1]-Max-Planck-Institute for Developmental Biology

johannes.soeding@tuebingen.mpg.de

HHpred3 is an automatic version of our structure and function prediction server HHpred (http://hhpred.tuebingen.mpg.de/) and is one of four related servers participating in CASP7 (HHpred1 to 3, BayesHH). It uses HMM-HMM comparison with integrated secondary structure comparison, correlation scoring, a novel local HMM-HMM maximum a-posteriori probability (MAP) alignment scheme, multiple template selection, and intermediate profile searching.

The tertiary structure prediction proceeds in five steps (all but step 3 are the same for HHpred2):

1. Build a multiple alignment from the target sequence with PSI-BLAST (1) (up to 8 rounds with E-value threshold 1E-3). PSIPRED (2) is used for secondary structure prediction.

2. The alignment is converted to an HMM and compared with a database of HMMs derived from representative sequences in the PDB, using the HHsearch software (3) in local Viterbi alignment mode.

3. If the top hit has a probability of less than 90% to be homologous, our intermediate profile search method HHsenser (4) is used to enrich the query alignment with more remote homologs.

4. The top 20 matches are clustered by UPGMA into a forest of separate trees, based on the structure comparison scores of TM-align (Zhang & Skolnick). The clustering stops when the highest average pairwise TM-score drops below 0.7. For each tree, a multiple structural alignment is calculated with MUSTANG (AS. Konagurthu et al.). The corresponding PSI-BLAST alignments are merged into a super-alignment in a master-slave fashion and an HMM is generated. The target HMM is compared with these HMMs and the best match defines a set of templates.

5. The top-scoring alignment with these templates is redetermined using the local MAP scheme.

6. MODELLER (A. Sali et al.) is used to generate a homology model from this multiple template alignment.

For function prediction, the target HMM from step 3 above is compared with the PDB and the Interpro database (8) using HHM-HMM comparison. Mappings to GO numbers are either provided by the GOA (9) and InterPro databases or, if these are not available, assigned by weighted word counts In this case, each word in the PFAM or PDB name and description text casts votes for GO terms containing this word. Words are wwighted depending on their frequency in the GO definition file and on a word frequency table for standard english.

For homology-based domain prediction, the target HMM from step 3 above is compared with the SCOP (10) and Pfam (11) databases using HHsearch in local Viterbi mode. The top alignments are realigned in global Viterbi mode and the aligned regions of the top-scoring hits overlapping not more than 20 residues and possessing at least 50 aligned residues define the domain boundaries.

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

2. Jones D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol. 292:195-202.
3. Söding J. (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics. 21:951-960.
4. Söding J., Remmert M., Biegert A., Lupas A.N. HHsenser: exhaustive transitive profile search using HMM-HMM comparison. Nucleic Acids Res. 2006 34:W374-8.
5. Zhang Y., Skolnick J. (2004) Scoring function for automated assessment of protein structure template quality. Proteins. 57:702-710.
6. Konagurthu A.S., Whisstock J.C., Stuckey P.J., Lesk A.M. (2006) MUSTANG: a multiple structural alignment algorithm. Proteins. 64:559-574.
7. Sali A., Blundell T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol. 1993 234:779-815.
8. http://www.ebi.ac.uk/interpro/
9. http://www.ebi.ac.uk/GOA/
10. http://scop.mrc-lmb.cam.ac.uk/scop/
11. http://www.ebi.ac.uk/interpro/

HIT-ITNLP - 459 models for 95 3D targets

The FRPPSP fold recognition method

Qi-wen Dong, Xiao-long Wang, Lin Lei

School of computer science and technology, Harbin institute of technology, Harbin, Chain.

Qwdong, wangxl, linl@insun.hit.edu.cn

We developed a Fold Recognition method by combining Profile-profile alignment and Profile-level Statistical Potentials (FRPPSP). The profile-level statistical potentials are described in our previous study[1], which use the evolutionary information of profiles and provide better discriminatory ability than those at the residue level.

In this study, the profile-level statistical potentials integrate the three single-body potentials, that is, the Φ/Ψ dihedral angle, accessible surface and contact statistical potentials:

$$E(i) = E^t(i,\phi_i,\varphi_i) + w^f E^f(i,S_i) + w^c E^c(i,N_i) \quad (1)$$

where Et, Ef, Ec is the Φ/Ψ dihedral angle, accessible surface and contact statistical potentials respectively, i is the profile type at the i-th position of the sequence, wf and wc are the weights of accessible surface and contact statistical potentials.

The profile-profile alignment method used here is the PICASSO3 method[2], which gives the best results of fold recognition . The profile-profile score to align the position i of a sequence q and the position j of a template t is given by:

$$m_{ij} = -\sum_{k=1}^{20}\left[f_{ik}^q S_{jk}^t + S_{ik}^q f_{jk}^t\right] \quad (2)$$

where fikq, fikt, Sikq and Sikt are the frequencies and the position-specific score matrix (PSSM) scores of amino acid k at position i of a sequence q and position j of a template t, respectively.

The profile-profile alignment is combined with the knowledge-based score for threading. The total score is given by:

$$u^{total} = m_{ij} + w^s E_j(s_i) \quad (3)$$

where Ej(si) is the combined potentials score of the template at position j with the residue type (for residue-threading) or profile type (for profile-threading) si of the position i of the query sequence, ws is the weight factors for structure scores. The dynamic programming algorithm is employed to find the minimum of the total score of the sequence-template alignments.

All profiles are generated by running PSI-BLAST3 on the NRDB90 database from EBI[4]. The five most favorable templates and the corresponding alignments are inputted to MODELLER [5] to generated the 3D structure.

1. Dong Q.W., Wang X.l., Lin L. (2006) Novel knowledge-based mean force potential at the profile level. BMC Bioinformatics;7:324.
2. Mittelman D., Sadreyev R., Grishin N. (2003) Probabilistic scoring measures for profile-profile comparison yield more accurate short seed alignments. Bioinformatics;19(12):1531-1539.
3. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J.H., Zhang Z., Miller W., Lipman D.J. (1997) Gapped Blast and Psi-blast: a new generation of protein database search programs. Nucleic Acids Research; 25(17):3389-3402.
4. Holm L., Sander C. (1998) Removing near-neighbour redundancy from large protein sequence collections. Bioinformatics; 14(5):423-429.
5. Sali A., Blundell T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol ; 234(3):779-815.

honiglab - 174 models for 90 3D/2 TR targets

Template-based protein structure prediction using an automated modeling pipeline, manual target analysis and fast model evaluation

D. Petrey[1], C.L. Tang[2], J. Zhu[1], A. Kuziemko[2], P. Liu[2], M. Shirts[3], S. West[2], L. Xie[1], S. Zhao[3], K. Zhu[3], R. A. Friesner[3] and B. Honig[1]

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

The central problem of template based protein structure prediction is, of course, identifying those regions of a target structure that have a conformation different from the equivalent region of its template(s), and modeling these regions using some *ab initio* method or through composite model building. Such regions can frequently be identified and modeled through an analysis the structural neighbors of potential templates and many successful methods use such an approach. As more and more experimentally determined protein structures become available, such approaches will concurrently grow in applicability. However, the optimal method of analyzing and using the structural neighbors is unclear, and the methods used at each stage of the modeling process can depend on the particular group of templates being considered. Consequently, a certain degree of expert analysis is often necessary.

We have developed an automated homology modeling pipeline whose design was guided by several principles: it should allow the use of a wide variety of methods depending on the characteristics of the target under consideration; it should allow the convenient examination and analysis of the templates, alignments and models at each stage of the prediction process; and it should allow a user to resubmit templates, alignments or models to the pipeline depending on the results of this analysis.

This pipeline was used extensively by our group during our participation in CASP7. A structure prediction for a given target was performed as follows. Our modeling pipeline was used to generate an initial set of models (usually ~100) in a completely automated way using a "standard" set of methods. In particular, template selection and alignment was carried out using a combination of our in-house profile-profile alignment tool HMAP[1] and the SP3[2] fold recognition tool. Models were constructed using NEST[3] and were evaluated using the statistical potential DFIRE[4] and Verify3D.[5] The interface to our pipeline was then used to analyze the automated modeling results. Several types of analysis can be carried out using this interface, including comparison of alignments of the target to different templates, querying of a functional annotation database, and creation of "project files" that can be viewed in the program GRASP2[6], which allows visualization and structure-based alignment of the templates and models, as well as manual alignment adjustment.

Based on the results of this analysis, selected templates/models were subjected to more comprehensive modeling. This decision was made based on several criteria: 1) The top ranked templates (in terms of statistical significance of the alignments) were checked for consistency in their function annotation. This was also carried out for the top-ranked models (based on the effective energy functions used). 2) As much as possible, where there were variations in the alignments of the target to different templates, all alignments were tried on all templates and evaluated based on the energy of the models. 3) Depending on the known characteristics of the target, primarily the presence of ligands and quaternary structure, templates that matched those characteristics were selected and models were generated that included the appropriate ligands and multimeric partners. 4) In situations where no single template had secondary structure that completely matched the predicted secondary structure of the target, models based on different templates were combined based on how well they locally matched the secondary structure prediction of the target.

Refinement was carried out primarily with two programs. Our in-house refinement method IMO and the program PLOP. IMO[7] uses a torsion-space local sampling algorithm, DFIRE and energy-driven clustering of the models. It was used for several purposes: to refine secondary structure elements in situations where the predicted length or type of secondary structure differed from the template, to combine models and to refine N- and C-terminal tails. PLOP[8] uses torsion-space sampling combined with all-atom energy functions and fast screening and clustering techniques to reduce the set of possible conformations to a small number of candidates that are evaluated via an optimized minimization algorithm. PLOP was used primarily for loop prediction, especially in situations where it was necessary to simultaneously model a loop as well as its nearby environment.

A final decision as to which model to submit was based on a combination of manual analysis and evaluation of the models using the statistical potential DFIRE.

1. Tang C.L., Xie L., Koh I.Y.Y., Posy S., Alexov E. and Honig B. (2003) On the Role of Structural Information in Remote Homology Detection and Sequence Alignment: New Methods Using Hybrid Sequence Profiles. J. Mol. Biol. . 334, 1043-1062.
2. Zhou H. and Zhou Y. (2005) Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins. 58, 321-328.

3. Petrey D., Xiang X., Tang C.L., Xie L., Gimpelev M., Mitors T., Soto C.S., Goldsmith-Fischman S ., Kernytsky A., Schlessinger A., Koh I.Y.Y., Alexov E. and Honig B. (2003) Using Multiple Structure Alignments, Fast Model Building, and Energetic Analysis in Fold Recognition and Homology Modeling. Proteins. 53, 430-435.
4. Zhou H. and Zhou Y. (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Prot. Sci. 11, 2714-2726.
5. Eisenberg D., Luthy R., and Bowie J. U. (1997) Verify3D: assessment of protein models with three dimensional profiles. Meth. Enz. 277, 396-404.
6. Petrey D, Honig B. (2003) GRASP2: Visualization, Surface Properties, and Electrostatics of Macromolecular Structures and Sequences. Meth. Enz. 374, 492-509.
7. Zhu J., Xie L. and Honig B. (2006) Structural Refinement of Protein Segments Containing Secondary Structure Elements: Local Sampling, Knowledge-Based Potentials and Clustering. Proteins. 65. 463-479.
8. Zhu K., Pincus D.L., Zhao S. and Friesner R.A. (2006) Long loop prediction using the protein local optimization program. Proteins. 65, 438-452.

hPredGrp- 100 models for 100 3D targets
Verify- 100 models for 100 3D targets

Fischerlab automatic predictions

T. Seth, M. Gattie and D. Fischer
{tseth, mtgattie, df33}@cse.buffalo.edu

We have submitted automated predictions using two approaches: metaprediction using the shub and beautshot servers and metaselection using the Verify, Taruna and MQAPcons procedures. Shub and beautshot are new autonomous servers generating models using an improved shotgun paradigm followed by the beautify refinement. Shub considers only models generated by inub, whereas beautshot includes also local implementations of the sp3 and prospect methods. To compare the value of the shotgun assembly, we submitted beautshotbase which corresponds to the same shotgun selection as beautshot but without the shotgun assembly. For metaselection we used models from about 10 servers and submitted the selected beautified model. Verify corresponds to Verify3D, Taruna is a shotgun-like selection (sum of maxsub all-vs-all comparisons, and without assembly) and MQAPcons is a combination of MQAPs with Taruna.

igor - 66 models for 52 3D targets

A Simple Easily-Integratable Model of a Protein Chain

M.J.Dudek
mdudek@nethere.com

A residue ising model is used as a simplified representation of a single polypeptide chain. We work with a 3 state model, meaning each residue exists in 1 of 3 possible secondary structure states: H=α-helix, E=extended strand of a β-sheet, or C=coil. The abbreviation ss will be used for secondary structure. Let nR be the number of residues in a chain. The states of a residue ising model consist of all mappings of the set of residues into the set of ss states {1,...,nR}-->{H,E,C}. The function that maps states to energies is formed as a sum of contributions each of which depends only on the ss state within a small window of 6 or less residues. These energy contributions, referred to as residue impulses, are attributed to patterns in the residue sequences within the small windows.

We define a ss element as a contiguous block of residues such that each residue of the block has the same ss state and any residue adjacent to the block on either side has a different ss state. The space of discrete chemical compositions for a ss element is the combinatorially large set of all possible lengths and residue sequences. We partition this space into 16 element types based on element length and 3 hydrophobic moments. An element composition for the chain is defined to be a partitioning of the chain into elements. There exist 2**(nR-1) possible element compositions corresponding to all possible choices of a set of element boundaries, or alternatively a binary yes or no choice of an element boundary at each residue boundary. Analogous to ising model representation of a residue sequence, we use also an ising model to represent the element sequence of an element composition. Let nG be the number of elements in a chain. While the set of residues is a fixed property of a chain, the set of elements differs for each choice of element composition. The states of the element ising model consist of all mappings of the set of elements

into the set of ss states {1,...,nG}-->{H,E,C} such that no 2 adjacent elements have the same ss state. The number of states, therefore, is 3*(2**(nG-1)). The function that maps states to energies is formed as a sum of contributions each of which depends only on the ss state within a small window of 5 or less elements. These energy contributions, referred to as element impulses, are attributed to patterns in the element sequences within the small windows.

In this work, we introduce a new model for representing a protein chain, which we refer to as an igor model. The ss states of an igor model are the same as the

3**(nR) ss states of a residue ising model. While a residue ising model is short range in nature, the more flexible form of the igor model energy function

attempts to account for medium and long range residue-residue interactions. Igor model energy is a function of element composition, a collection of 3*(2**(nG-1)) states, as opposed to a single whole chain ss state.

Let Z[res+elem](α) be the partition function for the element ising model representation of the element composition α that includes both residue and element impulses. Similarly, let Z[elem](α) and S[elem](α) be the partition function and entropy for the element ising model representation that includes only element impulses. The full energy, ε(α)= ε0(α) +ε1(α) +ε2(α) +ε3(α), is a sum of 4 components.

ε0(α) is an energy attributed to formation of single elements and short strings of elements.

ε1(α)= ln(3*(2**(nG(α)-1)) *Z[res+elem](α)/Z[elem](α)).

ε2(α)=-S[elem](α).

ε{3}(α) is an energy attributed to threading of the element ising model into the fold of a best-fit template from the set of scop40 domains. This energy includes the short range energy of distortion required for the element ising model to adopt the ss state of the aligned template, and the long range element-element interaction energies obtained by packing of the ss elements of the target in the fold of the template. We note that the level of crudeness of the element ising model is such that our use of alignment to templates is much more

a mechanism for sampling a large number of packing arrangements of ss elements, as opposed to a sensitive mechanism for fold recognition.

The igor model enables calculation of single residue ss state probability distributions and, more importantly, sampling of the collection of individual element compositions that contribute most to the partition function. Perhaps the major application for the igor model will be in protein structure prediction, to enable efficient sampling of backbone conformations to provide starting points for optimization of a more accurate energy function. A procedure for translating element compositions of the igor model into consistent full-atom structures uses homology model building based on the alignment to the best hit template.

ISTZORAN - 99 models for 99 DR targets

Length-Dependent Prediction of Protein Intrinsic Disorder

K. Peng[1], P. Radivojac[2], S. Vucetic[1], A.K. Dunker[3], Z. Obradovic[1]

[1] - Center for Information Science and Technology, Temple University, Philadelphia, PA 19122
[2] - School of Informatics, Indiana University, Bloomington, IN 47408
[3] - Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN 46202
zoran@temple.edu

The VSL2[1] predictor is a slightly improved version of the original VSL1[2] predictor which is applicable to disordered regions of arbitrary length. It consists of two specialized predictors, VSL2-L for *long* (>30 residues) and VSL2-S for *short* (≤30 residues) disordered regions, and a *meta* predictor to integrate outputs of the two specialized predictors. The final prediction is calculated as $O_L \times O_M + O_S \times (1 - O_M)$, where O_L, O_S, and O_M are outputs of VSL2-L, VSL2-S, and the meta predictor, respectively.

The training data for VSL2 consisted of 1,327 non-redundant protein sequences with pairwise identity ≤25%. In total there were 1,389 short and 217 long disordered regions with 34,911 residues. 483 very short disordered regions of 1-3 residues were not used in predictor training. The data also contained 406,342 ordered residues, about 8% of which came from regions of high B-factors. These residues were excluded since high B-factor regions are known to be similar to short disordered regions.

For all three component predictors, a same set of 51 features were constructed for each residue using a sliding window. These features included local amino acid frequencies, local sequence complexity, average net charge, average flexibility, average hydropathy, charge/hydropathy ratio, average PSI-BLAST profiles, average secondary structure predictions, and an additional one to indicate if the current residue is close to a terminus. The window lengths were chosen as 41 for VSL2-L, 15 for VSL2-S, and 61 for the meta predictor.

All component predictors were built as linear support vector machines (SVM) instead of the logistic regression models for VSL1. The SVM outputs were calibrated into posterior probabilities using a single-input logistic regression model. As in our previous studies, moving-average was applied to smooth the raw predictions to remove occasional misclassifications. The sliding window lengths for VSL2-L and VSL2-S, and meta predictor were 31, 5, and 1, respectively.

1. Peng K., Radivojac P., Vucetic S., Dunker A.K. & Obradovic Z. (2006) Length-dependent prediction of protein intrinsic disorder. BMC Bioinformatics 7, 208.
2. Obradovic Z., Peng K., Vucetic S., Radivojac P. & Dunker A.K. (2005) Exploiting heterogeneous sequence properties improves prediction of protein disorder. Proteins 61(S7), 176-182.

IUB-Info - 197 models for 81 FN targets

Protein function prediction from sequence, properties and literature

Wyatt T. Clark[1], Andreas Rechtsteiner[2], Amrita Mohan[1], Predrag Radivojac[1]

1 – School of Informatics, Indiana University, Bloomington, IN 47408
2 – Center for Genomics and Bioinformatics, Department of Biology, Indiana University, Bloomington, IN 47405
predrag@indiana.edu

Despite numerous high-throughput experimental efforts, the number of sequenced and translated proteins with unknown cellular function is growing rapidly. Currently, there are 385 completely finished genomes, while 56 archaeal, 933 bacterial, and 608 eukaryotic genomes are being sequenced around the world (http://www.genomesonline.org/). The major protein sequence repository, GenBank, contains over 3.6 million sequences, which includes all publicly sequenced/translated proteins as well as numerous isoforms and engineered sequences. At the same time, the number of non-redundant (<90% sequence identity) proteins with high-confidence annotations in the Gene Ontology database [1] is less than 25,000. Consequently, one of the major objectives of bioinformatics is to develop methodologies and tools for automated protein annotation that can be used by researchers working on individual proteins but also on a genomic scale. Here we report on a machine learning approach we used in CASP7 for the prediction in the "function" category.

Problem formulation. Given a query protein p, a set of functional terms $G = \{g_1, g_2, \ldots g_{|G|}\}$, and a set of proteins annotated with the terms from G, the goal is to output the subset of most likely annotation terms that characterize p. The outputs are sorted according to the approximated posterior probabilities for each particular $g \in G$.

Methodology. Data representation. Three groups of features were explored: (i) sequence alignment-based features, (ii) features based on amino acid sequence and it properties, and (iii) features based on literature. We note that property-based features were derived from protein amino acid sequence, but were separated because they reflect their physicochemical properties. High-dimensional data representation was developed for each functional term containing at least 10 non-redundant sequences. *Data selection and preprocessing.* To gather unbiased and high-quality data we eliminated sequence redundancy and used only functional evidence of high confidence. In particular, our starting dataset was an intersection of all proteins with appropriate GO evidence and the UniRef90 database. Only sequences with IEA IPI, IGI, TAS, IMP, IDA evidence codes were used as these annotations are most reliable. *Dimensionality reduction.* We explored feature selection filters to eliminate unpromising features and principal component analysis to remove feature correlation. *Model selection.* Linear support-vector machines were used [2]. One-versus-all training was performed and final outputs were combined from all individual models.

In CASP7 we evaluated 3 models: (i) model using sequence- , property- and literature-based features; (ii) model using sequence- and property-based features only, and (iii) model using literature-based features only.

1. Ashburner M., Ball C.A., Blake J.A., Botstein D., Butler H., Cherry J.M., Davis A.P., Dolinski K., Dwight S.S., Eppig J.T., Harris M.A., Hill D.P., Issel-Tarver L., Kasarskis A., Lewis S., Matese J.C., Richardson J.E., Ringwald M., Rubin G.M., and Sherlock G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25(1), 25-29.
2. Joachims T. (2002) Learning to classify text using support vector machines: methods, theory, and algorithms. Kluwer Academic Publishers.

JIVE - 458 models for 94 3D/7 TR targets

Prediction of protein structure using template-free assembly of secondary and super-secondary motifs.

David F Burke[1] and Tom L Blundell[1]

1 - Biochemistry Dept , University of Cambridge, 80 Tennis Court Road, Cambridge CB2 1GA
dave@cryst.bioc.cam.ac.uk

JIVE assumes a hierarchy of protein structure: amino acid conformation, secondary structure, super-secondary motif and globular subunit or domain. Predictions or knowledge of the secondary structure are used to influence predictions of super-secondary structure, based on conformational class predictions from the *SLoop* database[1,2]. These super-secondary structure predictions are combined to build up structures of larger modules and domains using a Monte Carlo simulation incorporating an A* algorithm[3,4] and a stochastic refinement protocol to remove local atomic clashes. These are then evaluated using filters describing known features of protein structure.

In CASP7, the secondary structure for each target was predicted using psipred[5], phd[6] and jpred[7] and a consensus produced. All of these were collectively used to predict super-secondary fragments of loops from the SLoop database. For comparative modelling targets, tertiary structure fragments were also derived from filtered models produced by CASP7 server predictions. Final models were assessed visually and models close to those produced by CASP7 servers were not submitted.

1. Burke D.F., Deane C.M. and Blundell T.L. (2000) A browsable and searchable web interface to the SLoop database of structurally-classified loops connecting elements of protein secondary structure. Bioinformatics: 16(6),513-519
2. Burke D.F. and Deane C.M. (2001) Improved Loop prediction from sequence alone Protein Engineering 14(7), 473-478
3. Hart P.E., Nilsson N.J. and Raphael B. (1968) A formal basis for the heuristic determination of minimum cost paths IEEE Transactions On Systems and Cybernetics SSC, 4(2), 100-107
4. Lermen M. and Reinert K. (2000) The Practical Use of the A* algorithm for exact multiple sequence alignments J.Comp.Biology, 7(5), 655-671
5. McGuffin L.J., Bryson K., Jones D.T. (2000) The PSIPRED protein structure prediction server. Bioinformatics 16, 404-405
6. Rost B., Sander C., Schneider R. (1994) PHD--an automatic mail server for protein secondary structure prediction. Comput Appl Biosci 10(1), 53-60
7. Cuff J.A., Clamp M.E., Siddiqui A.S., Finlay M., Barton G.J. (1998) JPred: a consensus secondary structure prediction server. Bioinformatics. 14(10) 892-3.

Jones-UCL - 322 models for 99 3D/83 QA/2 TR targets

Use of Fragment Assembly, Threading and Model Quality Assessment Methods to Predict Protein Folds

K. Bryson[1], M.I. Sadowski[1], A. Lobley[1] , S. Lise[1], C.S. Pettitt[1], L.J. McGuffin[2] & D. T. Jones[1]

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

[2] – The BioCentre, University of Reading, Whiteknights, PO Box 221, Reading RG6 6AS, UK

dtj@cs.ucl.ac.uk

Our pair potential based threading program THREADER[1] was used to predict targets which were not predicted with high confidence by mGenTHREADER[2]. However, in making full CASP7 submissions, we also considered other models obtained from our web servers along with functional information where available and the results from external servers. Typically, easier fold recognition and comparative modeling targets were built using a consensus approach from the top scoring mGenTHREADER hits. A simple C-alpha-based model refinement program (HOOKEMODEL) was used to splice the best models together and then fill-in any remaining gaps in the hybrid structure.

For CASP7 targets which we believed could not be reliably predicted using fold recognition methods, FRAGFOLD [3] was used to generate up to 5 structures. This approach to protein tertiary structure prediction is based on the assembly of recognized supersecondary structural fragments taken from highly resolved protein structures using a simulated annealing algorithm. The main changes to FRAGFOLD since CASP6 have been to a) improve the rotamer library used to build side chain positions, b) improve the steric energy function, c) improve the hydrogen bond function and d) greatly increase the efficiency of the program when handling multiple sequences. This latter feature is a major distinguishing feature of FRAGFOLD in that for family of sequences, each sequence is effectively folded in parallel and the energies for each sequence averaged. Between 1000 and 3000 structures were generated for each target using a 300-CPU Beowulf cluster, and a simple rigid-body structural clustering algorithm used to select the models representing the largest clusters of conformations. Submitted predictions were made using little or no human intervention apart from initial domain assignment and preparation of input secondary structure and sequence alignment files.

For all targets (including CM and FR targets), regions of native disorder were predicted using DISOPRED2 [4-5]. DISOPRED2 is based on a reimplementation of DISOPRED using Support Vector Machines rather than neural networks.

Several new Model Quality Assessment programs were tried in CASP7 alongside our existing MODCHECK[8] method. The most sophisticated of these is MODCHECK-EEKS (Everything Except the Kitchen Sink) which combines a very wide array of features in order to assess model quality. Components include an atomic solvation potential, side chain and main chain torsion angles, pair potentials using MODCHECK and detailed hydrogen bonding analysis. This method was used in the QA section of CASP, the model refinement section and to select the best FRAGFOLD models from the largest clusters in some cases.

1. Jones D.T., Taylor W.R. & Thornton J.M. (1992) A new approach to protein fold recognition. Nature 358, 86-89.
2. McGuffin L. J., Smith R.T., Bryson K., Sorensen S.A. & Jones D.T. (2006) High throughput profile-profile based fold recognition for the entire Human proteome. BMC Bioinformatics, 7, 288
3. Jones D.T. (1997) Successful ab initio prediction of the tertiary structure of NK-Lysin using multiple sequences and recognized supersecondary structural motifs. PROTEINS. Suppl. 1, 185-191.
4. Jones D.T. & Ward J.J. (2003) Prediction of disordered regions in proteins from position specific score matrices. Proteins. S6, 573-578.
5. Ward J.J., Sodhi J.S., McGuffin L.J., Buxton B.F., Jones, D.T. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J. Mol. Biol., 337, 635-645.
6. Pettitt C.S., McGuffin L.J. & Jones D.T. (2005) Improving sequence-based fold recognition by using 3D model quality assessment. Bioinformatics 21(17):3509-3515.

Jones-UCL (Servers) - 322 models for 99 3D/83 QA/2 TR targets

Protein Structure Prediction Servers at UCL

K. Bryson[1], L.J. McGuffin[2] & D.T.Jones[1]

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

[2] – The BioCentre, University of Reading, Whiteknights, PO Box 221, Reading RG6 6AS, UK

dtj@cs.ucl.ac.uk

In addition to our manual predictions for CASP7, we also entered predictions from a number of our publicly available servers. The first of these is a fold recognition server mGen3D, based on the mGenTHREADER[1-3] method. The core method has been improved through the use of a better profile-profile alignment algorithm[3] since CASP6. All of the parameters for the method have also been tuned using a genetic algorithm which optimized model quality over a set of 50 hard fold recognition targets.

The major new feature being tested in CASP7 is the generation of 3-D models from the fold recognition hits from mGenTHREADER. A simple algorithm is employed which potentially can take into account both continuous and discontinuous domains in the target sequence. In the first step, the top hit from mGenTHREADER is used to generate a main chain plus beta-carbon model. Subsequent alignments are then evaluated to see if they overlap with the first model. If a sufficient number of residues in a lower scoring alignment do not overlap with a region which has already been modelled then a new model is generated and evaluated for compactness. If a lower scoring alignment corresponds to a real domain then the new model should be approximately globular and compact. This process is continued until there are either less than 30 residues remaining or until the top 50 hits have been considered.

The DomPred Server[4] contains our previously published method for domain prediction, DomSSEA[5], combined with a newly developed method called Domains Predicted from Sequence (DPS).

DomSSEA uses a fold recognition approach, based on aligning the PSIPRED[6] predicted secondary structure for the query sequence against the observed secondary structures in a fold library. It then transfers the assigned domain boundaries from the best fold match to the query sequence.

DPS carries out a PSI-BLAST[9] search of the query sequence against a sequence database. Significant local alignment fragments are examined, and the total numbers of C- and N-terminals for the fragments are recorded for each residue position in the query sequence. These distributions are smoothed. They are then combined giving additional weight to positions which have high values for both the C- and N-terminals, since this provides more evidence for a domain boundary in which one conserved sequence region ends and another starts. The combined values are then turned into Z-scores by dividing throughout by the standard deviation over the entire query protein. A threshold is then applied to these z-score values in order to predict domain boundaries.

Lastly, the DISOPRED server[8] was evaluated. This server predicts regions of native disorder from sequence profiles using a Support Vector Machine.

1. Jones D.T. (1999) GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. J. Mol. Biol. 287, 797-815.
2. McGuffin L.J. & Jones D.T. (2003) Improvement of the GenTHREADER method for genomic fold recognition. Bioinformatics, 19, 874-881.
3. McGuffin L.J., Smith R.T., Bryson K., Sorensen S.A., & Jones, D.T. (2006) High throughput profile-profile based fold recognition for the entire

Human proteome. BMC Bioinformatics, 7, 288.
4. Bryson K., McGuffin L.J., Marsden R.L., Ward J.J., Sodhi J.S. & Jones D.T. (2005) Protein structure prediction servers at University College London. Nucl. Acids Res. 33(Web Server issue):W36-38.
5. Marsden R.L., McGuffin L.J. & Jones D.T. (2002) Rapid protein domain assignment from amino acid sequence using predicted secondary structure. Proteins Sci. 11, 2814-2824.
6. Jones D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292, 195-202.
7. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W. & Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389-3402.
8. Ward J.J., McGuffin L.J., Bryson K., Buxton B.F. & Jones D.T. (2004) The DISOPRED server for the prediction of protein disorder. Bioinformatics, 20, 2138-2139.

karypis - 161 models for 83 3D targets

Protein Structure Prediction using learning based methods, fragment assembly and simple alignment techniques.

H. Rangwala[1], C. Kauffman[1], K. Deronne[1], and G. Karypis[1],
1– University of Minnesota, Twin Cities.
{rangwala, kauffman, deronne, karypis}@cs.umn.edu

Our group participated in CASP7 manually and with three automatic servers: karypis.srv, karypis.srv.2, and karypis.srv.4. All follow the same basic protocol which begins with the selection of possible templates for a given target using profile and secondary structure information. This is followed by comparative modeling and model selection. These steps are described in detail below.

Given a query protein sequence, we primarily used DOMPro[3] to identify the possible domain boundaries which are further verified and changed based on domain prediction results of several other methods. Each predicted domain of a target is treated separately for subsequent steps.

The strategy of karypis.srv for template selection is to select based on a local alignment between the target and potential template. The alignment program uses profile and secondary structure based (YASSPP[1]) scoring to generate the top ten templates for each target. We rely on karypis.srv to generate templates for karypis.srv.4 and the manual prediction.

Alternatively, server karypis.srv.2 classifies the target domain sequences into one of 945 fold classes derived from the SCOP database (Version 1.69). Proteins belonging to the same fold tend to share the same structures, but may not exhibit high sequence similarity. We use direct profile based kernel methods[2] where we build 945 one-versus-rest discriminatory support vector machine based classifiers. Based on the prediction of these classifiers we are able to classify the domains into one of the folds. The top three scoring folds are selected and we then use the alignment scheme of karypis.srv to select from within each fold the top ten templates giving a total of thirty templates. For efficiency, the whole process is parallelized across 40 processors of a Linux-cluster. We also tried several other methods for selecting the best possible fold for target sequences. These include classifying the targets into one of the 1538 superfamilies (remote homology detection) and coupling the prediction output of the superfamily and fold level classifiers using a set of novel multi-class classification schemes[5].

After the generation of templates, each is aligned against the target and MODELER is used by karypis.srv, karypis.srv.2, and manual prediction to generate structures. All servers use a similar alignment technique to generate a target-template correspondence. Both servers employ side-chain refinement using SCWRL. In our manual submission, we use hand-tuned multiple structure alignments of several templates (generated with MUSTANG[6]) as a guide for MODELER..

We select from amongst the generated structures using several criteria. For karypis.srv and manual submission, the energy-based DOPE score produced by MODELER determines the top models. ProQ, a neural network method for structure quality evaluation, is employed by karypis.srv.2 to select the top models for submission.

Rather than use, MODELER, karypis.srv.4 constructs a model by assembling fragments of known protein structures[7] for five templates. Fragment placement is based on optimizing the RMSD between the working structure and the template. Models are evaluated based on their GDT_TS to the template.

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

4. Wallner B. and Elofsson A. (2003). Can correct protein models be identified ? Protein Science. (5), 1073-1086.
5. Rangwala H. S. and Karypis G. (2006) Building Multiclass Classifiers for Remote Homology Detection and Fold Recognition. BMC Bioinformatics (under review).
6. Konagurthu A.S., Whisstock J.C., Stuckey P.J., Lesk A.M. (2006) MUSTANG: a multiple structural alignment algorithm. Proteins: Structure, Function and Bionformatics. 64(3):559-74
7. DeRonne K.W. and Karypis G. (2006) Effective Optimization Algorithms for Fragment-assembly based Protein Structure Prediction. In Proceedings of Computational Systems Bioinformatics Conference (CSB).

keasar - 573 models for 100 3D/84 QA/2 TR targets

Refinement of Fold Recognition Models with Cooperative Solvation Potentials

Nir Kalisman, Ohad Greenshpan, and Chen Keasar
Department of Computer Science, Ben-Gurion Universiry, Israel
keasar@cs.bgu.ac.il

In this round of CASP our group submitted models for all CASP targets. All-atom refinement of FR models was the method of choice for most targets, although semi-automatic modeling without template was used for 15 targets with low fold-recognition scores. Initial FR alignments were downloaded from the 3D-Jury server[1], and all further modeling was done within the framework of MESHI[2]. Two novel solvation terms, detailed and coarse, were used extensively throughout the modeling process and are briefly described in this abstract. They differ in their levels of atomic detail and were used accordingly for various degrees of homology. Before modeling, the targets were sorted into easy/hard categories according to the quality of their alignments and their compliance with the predicted secondary structure. Each category was processed differently.

Easy targets. All-atom models were trivially generated from the top 20 alignments of 3D-Jury. They were evaluated with a scoring function that combined the coarse solvation energy of each model with its original 3D-Jury rank. The models were ranked according to the new score, and the top 5 were selected for further refinement. Selected alignments were manually curated, so that gaps were removed from template secondary structures and the burial of polar side chains was minimized. An all-atom model was created from each curated alignments, and its side chains were modeled concurrently by our program SCMOD. Gaps were completed and the entire model was submitted to energy minimization with standard terms and the detailed solvation term. Usually a set of a few hundred models was generated by slightly perturbing each model prior to minimization. The final models were selected according to their final minimization score.

Hard targets. Parts of the harder targets for which suitable templates were available were processed as in the easy target procedure. Next, the missing or unreliable parts of the templates were modeled according to the predicted secondary structures. The unreliable parts were then randomly perturbed and subjected to minimization that included the coarse solvation term. Models for submission were selected from the resulting model set according to the final minimization energy.

Modeling without templates. The secondary structure prediction was manually checked for the existence of clear structural motifs such as tight antiparallel beta sheets or strand-helix-strand. These motifs were modeled first. Next, the motifs along with the rest of the sequence were perturbed relative to each other and minimized with the coarse solvation term. Models for submission were again selected according to their final energy.

Coarse solvation term. The motivation for this term was to uncouple the side chain conformation from the backbone structure prediction, within the framework of the all-atom model. To this end, all the side chains were modeled to their most frequent backbone dependent rotamer. The solvation energy of each residue was then evaluated as a non-linear function of the number of neighboring carbon atoms in its first hydration shell. Hydrogen bonding and other polar interactions that require detailed placement of the side chains were ignored. The resulting term therefore assess the solvation of a residue based only on its backbone coordinates, and is therefore fast to compute. Yet important information about the probable side chain location is also included. The solvation is also less sensitive to small deviations of the backbone from its native state because the side chain positions are only approximated.

Detailed solvation term. In this term the solvation of each atom is a non-linear function of the number of neighboring carbon atoms in its first hydration shell. Yet, unlike similar solvent exclusion models, the number of hydrogen bonds in which the atom participates is also taken into consideration. As a result, the solvation of a buried polar atom that participates in a hydrogen bond is similar to its exposed state. This term is useful when the hydrogen network of the protein is partially known, i.e. when the backbone position is close to its native state.

1. Ginalski et. al. (2003) 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 19, 1015-8.
2. Kalisman et. al. (2005) MESHI: A new library of Java classes for molecular modeling. Bioinformatics 21, 3931-2.

KIHARA_PFP - 99 models for 97 FN targets

Partially automated, comprehensive annotation with PFP

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

For manual function prediction in CASP7, we partially automated and built on our multi-dimensional approach from CASP6. The process of defining functions for uncharacterized protein targets involved four major stages: (1) automatically annotating the target sequence with GO terms by PFP[1] and determining likely functional sites using MINER[2], (2) searching the target sequence against functional databases, (3) manually building and refining data from these primary searches, and (4) assigning additional GO or E.C. definitions to the target sequence based on our predicted 3D models. This method was used to gather predictions for the GO Molecular Function, Biological Process, and Cellular Component categories as well as E.C. definitions when applicable.

PFP[1] is a sequence-based function prediction algorithm which predicts GO terms for a target sequence based on term frequency in PSI-BLAST[3] results and contextual term association in annotated sequence databases. PFP is also implemented as a fully automated server, which participated in the function prediction server category of CASP7 (see group PFP_HAWKINS). MINER[2] is a multiple sequence alignment-based method which predicts functional sites in a target sequence whose phylogenetic trees have the most similarity to that of the complete sequence. PROSITE[4], PRINTS[5] and Blocks[6] were used for functional motif searching; Pfam and Pfam-FS[7] were used to for family alignments; PSORT[8] was used for subcellular localization; and STRING[9] was used for additional functional associations in primary searches. Information in the KEGG Pathway database[10] and thorough literature searches were used to refine and build on the data gathered from primary searches in the cases where that data was not sufficient to make a reasonable prediction of GO categories. Using this method, reasonable predictions were made for each of the 100 valid protein targets in CASP7.

1. Hawkins T., Luban S. & Kihara D. (2006) Enhanced automated function prediction using distantly related sequences and contextual association by PFP. Protein Sci. 15, 1550-1556.
2. La D. & Livesay D.R. (2005) MINER: software for phylogenetic motif identification. Nucleic Acids Res. 33, W267-W270.
3. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W. & Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389-3402.
4. Sigrist C.J.A., Cerutti L., Hulo N., Gattiker A., Falquet L., Pagni M., Bairoch A. & Bucher P. (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. Brief Bioinform. 3, 265-274.
5. Attwood T.K., Bradley P., Flower D.R., Gaulton A., Maudling N., Mitchell A.L., Moulton G., Nordle A., Paine K., Taylor P., Uddin A. & Zygouri C. (2003) PRINTS and its automatic supplement, prePRINTS. Nucleic Acids Res. 31, 400-402.
6. Henikoff S., Henikoff J.G. & Pietrokovski S. (1999) Blocks: A non-redundant database of protein alignment blocks derived from multiple compilations. Bioinformatics. 15, 471-479.
7. Bateman A., Coin L., Durbin R., Finn R.D., Hollich V., Griffiths-Jones S., Khanna A., Marshall M., Moxon S., Sonnhammer E.L.L., Studholme D.J., Yeats C. & Eddy S.R. (2004) The Pfam Protein Families Database. Nucleic Acids Res. 32, D138-D141.
8. Nakai K. & Kanehisa M. (1991) Expert system for predicting protein localization sites in Gram-negative bacteria. PROTEINS: Structure, Function, and Genetics. 11, 95-110.
9. Mering C., Huynen M., Jaeggi D., Schmidt S., Bork P. & Snel B. (2003) STRING: a database of predicted functional associations between proteins. Nucleic Acids Res. 31, 258-261.
10. Kanehisa M. & Goto S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27-30.

KIST - 450 models for 97 3D targets

Simulation of Protein Folding Structures

Chang No Yoon[1], Myung Whan Chi[1], Jin Su Song[1], Young Sun Kim[1], and Jin Kak Lee[1,2]
[1] - Korea Institute of Science and Technology, Cheongryang, Seoul, Korea
[2] - Nanormics, Inc. 10-57 Hawolgokdong, Sungbukku, Seoul, Korea
cody@kist.re.kr

To simulate the folding structures of a protein, we used a simple off-lattice model with the unified-residue point, which represents the alpha carbon of each amino acid in the protein model. This model has two angle variables, one for the angle between two consecutive virtual bonds, residues i to j and j to k, the other for the rotational angle of the virtual bonds consisting of residues i, j, k and l. In order to generate the structural conformations the Monte Carlo method was used with the starting point of random coil conformations. During this

procedure the range of the i-j-k angle was limited between 60 to 150 degrees. Among the trajectory data obtained from the navigation through the potential surface, about half of them were accepted and stored. The knowledge-based potential was used to obtain the potential energy surface. It was derived from the known protein structures. The total number of the accepted conformations was about 10E3 and the total steps for one run were about 10E8. Finally, all the conformations were clustered using the energy and RMS between the alpha carbon traces. Then the obtained representative conformations were minimized with the potential energy.

kitaura-fams - 8 models for 8 3D targets

Refinement of protein structures using the fragment molecular orbital (FMO) method

Toyokazu Ishida[1], Dmitri G. Fedorov[1], K. Kanou[2],
M. Takeda-Shitaka[2], H. Umeyama[2], and Kazuo Kitaura[1]
[1] - National Institute of Advanced Industrial Science and TechnologyiAIST)
[2] - Department of Biomolecular Design School of Pharmacy, Kitasato University
kanouk@ pharm.kitasato-u.ac.jp (toyokazu.ishida@aist.go.jp)

The fragment-based *ab initio* MO method (the FMO method[2]) was applied to the refinement of the target protein structures, tr288, tr368 and tr370. In the method, a molecule is divided into fragments and *ab initio* MO calculations are performed on the fragments and their dimers to obtain the total energy and other properties of the whole molecule. The FMO method reproduces regular *ab initio* MO results with high accuracy, hence molecular geometry optimized with this method is expected to have nearly *ab initio* quality. The method has been incorporated into GAMESS program package[2] with an efficient parallel algorithm (GDDI[3]), which was used for all FMO calculations in this work.

In the geometry optimization calculations all degrees of freedom were optimized at the FMO-RHF/3-21G level of theory with one residue/fragment partition of the proteins except for Gly which grouped with its neighboring residue because of its small size. We prepared the initial geometry by adding missing hydrogen atoms to the given coordinate data assuming the standard charge state of residues; Glu and Asp were deprotonated and Arg, Lys and His protonated. Preceding the FMO calculations, a rough minimization (about 100 steps) was carried out with the Amber96 force field to remove unphysically short contacts of atoms in the given protein geometry.

Geometry optimization of tr288 was completed; the maximum gradient (MaxG) was less than the convergence criterion of $5*10^{-4}$ Hartree/Bohr. The computational time was 367 hours on 180 2.0 GHz Opteron CPUs. The geometry optimizations of tr368 and tr370 were not completed by the deadline and their final MaxGs were $1.5*10^{-2}$ and $4.4*10^{-3}$ Hartree/Bohr, respectively. The reported geometries of these proteins, therefore, are not final.

Geometry optimizations of gas-phase proteins often result in the proton-transfers along salt-bridges. In tr368 and tr370 proton transfers occurred along several salt-bridges: Arg19-Asp16, His41-Asp55, and His59-Asp55 (tr368), as well as Arg71-Gln86, Asp48-Arg124 and Asp154-Arg12 (tr370), and the protons migrated between the acidic and basic sites during the optimizations. According to our experience, the geometry distortion due to the proton-transfer is limited to local regions and global conformations are not changed greatly. The relatively small RMS gradient ($4.6*10^{-4}$ and $3.6*10^{-4}$ Hartree/Bohr, for tr368 and tr370, respectively) suggests that the large gradient values are limited to atoms involved in the proton-transfers and the gradients for the majority of atoms are rather small. So the refined geometries of tr368 and tr370 at the present level may be useful although their optimizations are not completed.

Because there was not enough time to perform the FMO calculation, the same method as fams-multi team (see fams-multi abstract) had been applied for the targets of tr322, tr362, tr367, and tr380. Fams-multi had participated in refinement experiment using Energy minimize & Molecular dynamics. Under some constraint conditions to maintain no great conformation-change, the refined models were correctly revised for hydrogen bonds, main-chain torsion angles, side-chain torsion angles and the decreasing collision between hydrophobic atoms.

1. Fedorov D.G. and Kitaura K (2006) Modern methods for theoretical physical chemistry and biopolymers, edited by E. Starikow, S. Tanaka and J. Lewis, Elsevier, Amsterdam, pp. 3-38.
2. GAMESS, http://www.msg.ameslab.gov/GAMESS/GAMESS.html
3. Fedorov D.G., Olson R.M., Kitaura K., Gordon M.S., Koseki S. (2004) J.Comp.Chem., 25, 872-880.

KORO - 200 models for 40 3D targets

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

T.N. Sasaki[1], K. Imai[2, 3], T. Tsuji[1], S. Mitaku[1], and M. Sasai[1].
1) Department of Computational Science and Engineering, Graduate School of Engineering, Nagoya University, Furocho, Chikusa, Nagoya 464-8603, Japan, 2) Toyota Physical and Chemical Research Institute Nagakute, Aichi 480-1192,Japan, 3) Department of Applied Physics, Graduate School of Engineering, Nagoya University, Furocho, Chikusa, Nagoya 464-8603, Japan
sasai@tbp.cse.nagoya-u.ac.jp

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

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

From these structures obtained from folding simulations we selected the model structures by using the energy criterion and the cluster analysis. For smaller targets, the model_1 and model_2 structures were the lowest and second-lowest energy structures, and the other 3 model structures were selected from the results of the clustering analysis. For larger targets, we mostly selected the 5 low-energy structures as the 5 model structures. Additionally, we sometimes used SOSUIbreaker[4, 5] to check the results.

1. Sasaki T.N., & Sasai M. (2005) A coarse-grained Langevin molecular dynamics approach to protein structure reproduction, Chem. Phys. Lett. 402, 102-106.
2. Ginalski K., Elofsson A., Fischer D., & Rychlewski L. (2003) 3D-Jury: a simple approach to improve protein structure predictions, Bioinformatics 22, 1015-1018.
3. Cheng J., & Baldi P. (2005) Three-stage prediction of protein beta-sheets by neural networks, alignments and graph algorithms, Bioinformatics 21, Suppl 1;i75-84.
4. Imai K., & Mitaku S. (2005) Mechanisms of secondary structure breakers in soluble proteins, BIOPHYSICS 1, 55-65.
5. Imai K., Asakawa N., Tsuji T., Sonoyama M. & Mitaku S. (2005) Secondary structure breakers and hairpin structures in myoglobin and homeoglobin, Chem-Bio Info. J. 5 65-77.

Largo - 28 models for 1 3D/27 QA targets

Quality Assessment of 3D-models by LIBRA_rotamer

K. Tomii[1] and M. Ota[2]
[1]-Computationa1 Biology Research Center, National Institute of Advanced Industrial Science and Technology, 2-42 Aomi, Koto-ku, Tokyo, Japan [2]-Global Scientific Information and Computing Center, Tokyo Institute of Technology, O-okayama, Meguro-ku, Tokyo 152-8550, Japan
k-tomii@aist.go.jp

In template-based modeling, and even in the case of template-free modeling, it is important to select more accurate 3D-models among a pool of candidate structures produced by a given appropriate way. However, our conventional method to select the best model for a target sequence does not always succeed. For that reason, a more effective scheme is required for 3D-model evaluation. To this end, we used the new evaluation function, the so-called LIBRA_rotamer.

The LIBRA_rotamer was originally developed for threading and protein sequence design. It checks side-chain packings, hydration, local conformations, and repulsions of 3D-models. The side-chain packing term is a function of amino acid pair types, spatial distances, and types of side-chain rotamers. A rotamer library including 56 templates was used. The side-chain packing function is defined when the sequence separation is greater than four residues.

The hydration function is defined by the number of surrounding heavy atoms. The local conformational classes are defined by conformations of penta-peptide fragments. The local conformational function is also defined for each rotamer (see ref.1 for details).

We recognized that the correlation between the accuracies of our submitted models in previous CASP and the scores of LIBRA_rotamer is better than those obtained using our conventional method. In this study, we tested the ranking ability of the function by assessing the quality of the models that were submitted by prediction servers.

We ranked all submitted 3D-models by prediction servers using this evaluation function, according to the assessment scheme in CASP7. Because the coordinates of sidechain atoms are necessary for evaluation using the function, only the structural quality scores for 3D-models that possess sidechains are calculated. The correlation between the model accuracy and scores becomes obscure at the bad (high) zone of scores. Therefore, we generally ranked 3D-models with good (low) scores only.

1. Ota M., Isogai Y. & Nishikawa K. (2001) Knowledge-based potential defined for a rotamer library to design protein sequences. Protein Eng. 14, 557-564.

LCBDavis - 91 models for 91 FN targets

Prediction of protein function using local descriptors of protein structure

T.R. Hvidsten[1], A. Kryshtafovych[2], P. Daniluk[2], K. Fidelis[2] and Jan Komorowski[2]

[1] - The Linnaeus Centre for Bioinformatics, Uppsala University, Uppsala, Sweden, [2] - UC Davis Genome Center, Davis, USA

hvidsten@lcb.uu.se

Although tools such as BLAST[1] revolutionized experimental biology by providing testable hypotheses of protein function, identifying functionally characterized homologues using sequence similarity is only possible for less than 50% of the proteins predicted from genome sequencing projects. Since structure is evolutionarily more conserved than sequence, it is believed that experimental structures and predicted models from structural genomics projects may provide a solution for the remaining proteins[2].

We have developed a new method for representing and comparing protein structure based on *local descriptors of protein structure*[3]. A local descriptor is a set of short backbone fragments centered in three dimensions around a particular amino acid. A local descriptor is built by a) identifying all close amino acids within a radius of 6.5 Å, b) for each close amino acid, adding four sequence neighbors, two from each side, to obtain continuous backbone fragments of five amino acids, and c) merging any overlapping fragments into segments. We first computed local descriptors from all amino acids in a representative set of protein domains from PDB with less than 40% sequence identity to each other (ASTRAL version 1.63[4]). We then constructed a library of commonly reoccurring local descriptors by a) for each local descriptor identifying the group of all structurally similar local descriptors and b) selecting a set of 3720 representative, partially overlapping *descriptor groups*.

We represented all protein structures in ASTRAL in terms of structurally matching or not matching each of the local substructures in the descriptor library. For CASP targets we matched the local descriptors to structures predicted by Robetta[5]. In addition to structure, we added sequence information in terms of PROSITE[6] patterns and matches to families in Pfam[7].

We used the ROSETTA system[8] to model the relationship between sequence/structure and function with IF-THEN rules. The rules consist of minimal combinations of properties (local substructures and/or sequence motifs/families) (IF-part) that discriminate one molecular function from other, discernible functions (THEN-part). Function predictions are obtained based on the combined evidence given by all matching rules.

The rule model was induced based on 3963 Gene Ontology (GO)[9] annotation, distributed over 87 molecular function classes, to 2541 proteins in ASTRAL. Using 10 fold cross validation we were able to correctly predict 68% of these annotation, and at least one correct prediction for 74% of the proteins, with 47% of the predictions being correct. For CASP, the GO predictions were also mapped to EC numbers using ec2go.

The approach described here represent a model-based approach to function prediction in which a general library of local substructures, capable of assembling large parts of most proteins in ASTRAL, are used to describe protein functions as given by a set of training examples.

1. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W. & Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389-3402.
2. Chandonia J.M. & Brenner S.E (2006) The impact of structural genomics: expectations and outcomes. Science 311, 347-51.
3. Hvidsten T.R., Kryshtafovych A., Komorowski J. & Fidelis K. (2003) A novel approach to fold recognition using sequence-derived properties from sets of structurally similar local fragments of proteins. Bioinformatics 19 Suppl 2, II81-II91.

4. Brenner S.E., Koehl P. & Levitt M. (2000) The ASTRAL compendium for protein structure and sequence analysis. Nucleic Acids Res 28, 254-6.
5. Kim D.E., Chivian D. & Baker D. (2004) Protein structure prediction and anlaysis using the Robetta server. Nucleic Acids Res. 32, W526-W531.
6. Hulo N., Bairoch A., Bulliard V., Cerutti L., De Castro E., Langendijk-Genevaux P.S., Pagni M. & Sigrist C.J.A. (2006) The PROSITE database. Nucleic Acids Res. 34, D227-D230.
7. Finn R.D., Mistry J., Schuster-Böckler B., Griffiths-Jones S., Hollich V., Lassmann T., Moxon S., Marshall M., Khanna A., Durbin R., Eddy S.R., Sonnhammer E.L.L. & Alex Bateman (2006) Nucleic Acids Res. 34, D247-D251.
8. Komorowski J., Øhrn A. & Skowron A. (2002) The ROSETTA Rough Set Software System. In Handbook of Data Mining and Knowledge Discovery (eds. Klösgen, W. & Zytkow, J.) 554-559 (Oxford University Press).
9. Ashburner M. et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25, 25-9.

LEE - 1098 models for 99 3D/99 DP/99 QA/9 TR targets

A Template based Modeling based on Global Optimization

Keehyoung Joo[1], Jinwoo Lee[1], Sunjoong Lee[2], Joohyun Seo[3], Sung Jong Lee[4], Jooyoung Lee[1]

[1]-School of Computational Science, Korea Institute for Advanced Study,
[2]-Department of Physics, Hanyang University, Korea,
[3]-School of Chemical and Biological Engineering, Seoul National University, Korea, [4]-Department of Physics, Suwon University, Korea
jlee@kias.re.kr

For the prediction of the 3D structures of 100 CASP7 targets, we have developed a procedure which is based on global optimization of score functions in three levels. The whole procedure is composed of the following five steps:

1. Fold recognition: To collect fold candidates of a given target sequence, we considered top scoring templates from the meta-server provided by http://bioinfo.pl/~3djury, and another top scoring templates from an in-house method called FoldFinder. FoldFinder is a profile-profile alignment method utilizing predicted secondary structures. We have used a fold database of 17930 protein chains obtained from PISCES [1] at the 99 % sequence identity level. After collecting these templates, we performed a preliminary assortment of structural clustering often leading to 2 or 3 sets of template lists. These lists are the input to the following procedure.

2. Multiple sequence/structure alignment by MSACSA: This is the most crucial and computationally time consuming part of the method. We have performed multiple sequence/structure alignment for each template list obtained from the fold recognition step. Unlike the other heuristic (progressive) alignment methods popular in the literature, we have applied a rigorous global optimization method to an in-house consistency-based scoring function similar to the COFFEE [2] by using the conformational space annealing [3] (CSA) method. We have constructed a pair-wise restraint library generated from profile-profile alignment between the query sequence and template sequences and structure-structure alignment between templates using TM-align [4]. The lowest scoring alignment among the 100 final ones from the CSA is used as the input to the following 3D modeling step. The maximum number of templates performed in the multiple alignment for this CASP7 was 25.

3. Modeling of the 3D structure by ModellerCSA: The 3D structures of target proteins are constructed by optimizing the MODELLER [5] energy function using the CSA method. For each multiple alignment (containing up to 25 templates), a total of 100 models are generated and they are used for the list selecting procedure in the following step. This is the second most computationally time consuming part of the method.

4. List selection and the clustering of models for final model selection: For most cases, we have more than one list of templates, and we have applied a neural network based in-house procedure to assess the quality of the models obtained for each list. From the dominating winning list (if exists), we have applied the clustering method SPICKER [6], to find the center model of the cluster. We also selected lowest scoring models in terms of the Modeller energy and/or DFIRE [7] energy. When there are competing lists, we have used more than one list to select 5 models for final submission.

5. Side-chain modeling for selected targets by ROTCSA: For targets that we have decided worth side-chain modeling, we have constructed side chains as follows. For each list, a rotamer library is constructed based on the consistency of the side chains in the final 100 models obtained in the step 3. To this library, we have added a backbone dependent and sequence specific rotamer library similar to the SCWRL3.0 [8]. Using the CSA, we have optimized an in-house scoring function which contains energy terms from SCWRL and DFIRE.

1. Wang G. and Dunbrack R. L., Jr., (2003) Bioinformatics, 19;1589-1591
2. Notredame C., Holme L. and Higgins D.G. (1998) Bioinformatics 14(5);407-422
3. Lee J., Scheraga H.A. and Rackovsky S. (1997) J. Comput. Chem. 18;1222–1232
4. Zhang Y. and Skolnick J. (2005) Nucleic Acids Research, 33;2302-2309
5. Sali A. and Blundell T.L. (1993) J. Mol. Biol. 234;779-815
6. Zhang Y. and Skolnick J. (2004) J. Comput. Chem. 25(6);865-871

7. Zhou H. and Zhou Y. (2002) Protein Science, 11;2714-2726
8. Canutescu A.A., Shelenkov A.A. and Dunbrack R.L., Jr. (2003) Protein Science, 12;2001-2014

Levitt - 8 models for 8 TR targets

Pairwise Atomic Potentials and Near-Native Structure Refinement: *In vacuo* energy minimization

C.M. Summa[1] and M. Levitt[1]
[1] –*Stanford University*
csumma@stanford.edu

Several molecular mechanics force fields are compared for their ability to attract a near-native decoy protein structure towards the native structure. This problem is closely linked to the techniques of protein homology modeling, structure prediction, and refinement. A dataset of 75 structurally diverse proteins was constructed, and for each of these proteins 729 near-native decoys were generated by perturbing the structure along its 6 lowest frequency non-orthogonal normal modes. We tested several traditional molecular mechanics potentials (AMBER99 [1, 2], GROMOS 43B1 [3], OPLS-AA [4], and ENCAD [5]) using a powerfully convergent energy minimization method and show that, of the traditional molecular mechanics potentials tested, only one, AMBER99, showed a modest net improvement in <cRMS> over the set of near native decoys. A smooth, differentiable knowledge-based pairwise atomic potential was also generated in the manner of Skolnick [6], and was shown to perform much better on this test than any of the traditional potential functions tested, including AMBER99.

1. Wang J., Cieplak P., and Kollman P.A. (2000) How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? J. Comput. Chem., 21(12): p. 1049-1074.
2. Sorin E.J. and Pande V.S. (2005) Exploring the helix-coil transition via all-atom equilibrium ensemble simulations. Biophys J. 88(4): p. 2472-93.
3. Van Gunsteren W.F., et al., Biomolecular Simulation: the GROMOS96 Manual and User Guide. 1996, Zurich, Switzerland: Vdf Hochschulverlag AG an der ETH Zürich.
4. Kaminski, G.A., et al., Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides. J. Phys. Chem. B, 2001. 105(28): p. 6474-6487.
5. Levitt, M., et al., Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. Comp. Phys. Comm., 1995. 91: p. 215-31.
6. DeBolt, S.E. and J. Skolnick, Evaluation of atomic level mean force potentials via inverse folding and inverse refinement of protein structures: atomic burial position and pairwise non-bonded interactions. Protein Engineering, 1996. 9(8): p. 637-55.

Ligand-Circle - 439 models for 88 3D targets

Ligand-Circle: CIRCLE to evaluate binding sites

Daisuke Takaya, Genki Terashi, Mayuko Takeda-Shitaka, Kazuhiko Kanou, Mitsuo Iwadate, Akio Hosoi, Kazuhiro Ohta and Hideaki Umeyama
Department of Biomolecular Design, School of Pharmacy,Kitasato University
p99150@st.pharm.kitasato-u.ac.jp

We have developed CIRCLE[1] since previous CASP because we didn't have high-precision scoring function for tertiary structure. CIRCLE has capability of evaluating tertiary structure with ligand optionally. Ligand-Circle aims for selecting the best model having high accuracy binding site in server models (TS+AL). So this team participated in TS category in CASP7.

Collecting server models

Server models were obtained from CASP7 home page (http://www2.predictioncenter.org/index_serv.html).

Generating refined model from servers

These models include tertiary structure (TS) and alignment (AL), and therefore these were refined or changed to tertiary structure by FAMS[2]. If it was AL format, a model was built based on this alignment. If it was TS format, a model is refined by FAMS. We used all the server models as its template because these models include CA model or having lacking residue. Moreover, our CIRCLE 3D1D method needs side chain coordinates [1].

Superimposing and evaluating

Experimentally known structures having ligand were obtained and superimposed to a refined server model using CE program[3]. The list of superimposed PDBID was gotten from PARENT of server. PDBID not having ligand was ignored.

Ranking refined models.

CILCLE score with ligand was calculated for these refined model with ligand , and we ranked the order using this score.

Result

server name	CHI1	**server name**	**chi1**
ROBETTA_TS1	5039	HHpred1_TS1	4163
Pmodeller6_TS1	4995	HHpred3_TS1	4160
Pcons6_TS1	4847	3D-JIGSAW_POPULUS_TS1	4139
FAMSD_TS1	4813	FOLDpro_TS1	4136
FAMS_TS1	4755	beautshotbase_TS1	4124
Zhang-Server_TS1	4700	BayesHH_TS1	4120
CIRCLE_TS1	4681	shub_TS1	4106
Ligand_Circle_TS1	4580	MetaTasser_TS1	4044
FUNCTION_TS1	4576	RAPTORESS_TS1	4038
CaspIta-FOX_TS1	4536	karypis.srv_TS1	4038
SAM_T06_server_TS1	4532	RAPTOR_TS1	4026
Bilab-ENABLE_TS1	4498	keasar-server_TS1	4006
Phyre-1_TS1	4487	SP4_TS1	3992
Phyre-2_TS1	4457	SP3_TS1	3991
PROTINFO_TS1	4360	RAPTOR-ACE_TS1	3981
HHpred2_TS1	4229	SPARKS2_TS1	3960
beautshot_TS1	4188	LOOPP_TS1	3947
3Dpro_TS1	4172	3D-JIGSAW_RECOM_TS1	3913

(5 targets which Ligand-Circle can't submit were not included)

This table shows $\chi 1$ ranking of 76 server models at 2 Oct. 2006. "chi1" means the number of correct $\chi 1$[4] We concentrated in the correctness of $\chi 1$ angle since the correctness in the binding site will depend upon that of $\chi 1$ angle of the side chain. Ligand-Circle is efficient.

1. See "CIRCLE: Full automated homology-modeling server using the 3D1D scoring functions" item in this book.
2. Ogata K. and Umeyama H. (2000) J. Mol. Graphics Mod. 18 258-272.
3. Shindyalov I.N., Bourne P.E. (1998) Protein Engineering 11(9) 739-747.
4. Fischer D., Elofsson A., Rychlewski L., Pazos F., Valencia A., Rost B., Angel R. Ortiz, and Dunbrack R.L. Jr. (2001) Proteins 5 171–183

LMU - 191 models for 65 3D/98 DP targets

Fold recognition and Alignment optimization using

AutoSCOPe and protein class flexibility

Fabian Birzele, Gergely Csaba, Ralf Zimmer

Ludwig-Maximilians-University Munich,
Amalienstr. 17, 80333 München, Germany
{Fabian.Birzele|Gergely.Csaba|Ralf.Zimmer}@bio.ifi.lmu.de

For CASP7 we tried a combination of server prediction and manual evaluation and adjustment of alignment. First (Phase I), we applied a pipeline of programs to each target and, second (Phase II), we analysed the results based on a database of multiple structural alignments of all pdb protein domains in order to decide on known or new fold, superfamily and family. We submitted only targets we think have a known fold and adjusted the computed alignments based on the flexibility observed in the multiple structure alignments in our database.

This database [Csaba, 2006] of multiple annotated structural alignments has been determined based on a new measure of structural similarity, which tries to account for some structural flexibility in protein structures. In addition to optimizing structural criteria such as RMSD and TM-score the alignments also try to conserve important functional and interaction sites of the proteins of the respective class.

In Phase I, we applied SSEP-Domain [Gewehr et al., 2006] to determine possible structural domains of the target protein and applied the following steps to each of the detected domains separately. Then we used PsiBlast to identify clear homologues, our new approach AutoSCOPe [Gewehr et al., 2006b] to identify known families, superfamilies, or folds based on so called 'unique patterns' detected in the target sequence. In addition, we tried a fold and family recognition with several alignment methods: Profile-Profile-Alignment (PPA) [von Oehsen et al., 2001-2005] of PsiBlast profiles for target and template with additional secondary structure information, SSE-align [Gewehr et al, 2006] a method matching predicted and actual secondary structure elements, and the quite old 123D threading method [Alexandrov et al., 1996] enhanced with profile information of target and template, or both, as well as secondary

structure information (123D+). For the alignments, we did not exploit knowledge or parameters from multiple structure alignments.

In Phase II, we manually analysed the results of PSiBlast, AutoSCOPe, SSE-align, PPA, and 123D in order to identify a consensus of fold, superfamily, or family and to select the best template. For this we also analysed the respective alignments, i.e. we computed alignments for all possible templates with SSE-align, PPA and 123D+. The alignments were evaluated with QUASAR [Birzele et al, 2005] checked with respect to coincidence of predicted and template secondary structures and most important the fit of features of the target sequence with features of the template class as derived from the multiple structure database [Csaba, 2006; Gewehr et al., 2006a]. In a couple of cases we manually adjusted the alignment to make it compatible with the predicted functional sites known to be conserved in the multiple alignment of the template. In addition, we used Vorolign [Birzele et al, 2006], a new structural superposition method for identifying structurally similar proteins based on Voronoi decompositions of the structures. Vorolign helped to find structurally similar proteins for a candidate template and to judge conserved and flexible parts of the template structure. Based on the multiple alignments and Vorolign we often de-aligned parts of the target sequence in order to account for the predicted structural flexibility in the target protein.

1. Alexandrov N.N., Nussinov R. and Zimmer R.M. (1996). "Fast protein fold recognition via sequence to structure alignment and contact capacity potentials." Pac Symp Biocomput: 53-72.
2. Birzele F., Gewehr J.E. and Zimmer R. (2005) QUASAR - Scoring and Ranking of Sequence-Structure Alignments. Bioinformatics, Vol. 21, No. 24, 2005, 4425-4426.
3. Birzele F., Gewehr J.E., Csaba G. and Zimmer R. (2006) Vorolign - Fast Structural Alignment using Voronoi Contacts. Accepted for ECCB 2006, to appear in Bioinformatics.
4. Csaba G. (2006). Analysis of Protein Sequence-Structure Alignments. Department of Informatics. LMU München.
5. Gewehr J.E. and Zimmer R. (2006) "SSEP-Domain: protein domain prediction by alignment of secondary structure elements and profiles." Bioinformatics 22(2): 181-187.
6. Gewehr J.E., Szugat M., Macri A. and Zimmer R. (2006) ProML: Flexible description of proteins and protein sets based on Biotypes. submitted
7. Gewehr J.E., Hintermair V. and Zimmer R. (2006) AutoSCOPe: Automated Prediction and Characterisation of SCOP Classes using Sequence Patterns. submitted
8. von Öhsen N. and Zimmer R. (2001) Improving profile-profile alignments via log average scoring. Algorithms in Bioinformatics (WABI 2001), Aarhus, Springer.
9. von Öhsen N., Sommer I. and Zimmer R. (2003) Profile-profile alignment: a powerful tool for protein structure prediction. Pac Symp Biocomput: 252-63.
10. von Öhsen N., Sommer I., Zimmer R. and Lengauer T. (2004) Arby: automatic protein structure prediction using profile-profile alignment and confidence measures. Bioinformatics 20(14): 2228-35.
11. von Öhsen N. (2005) Profile-Profile alignment for remote homology and domain detection of proteins. Computer Science. Munich, Ludwig-Maximilians-University.

LOOPP - 500 models for 100 3D targets

Learning, Observing and Outputting Protein Patterns (LOOPP)

Jaroslaw Pillardy[1], Jian Qiu[2], Brinda K. Vallat[3], Ron Elber[4]

1 - Computational Biology Service Unit, Cornell University
2 - Cornell University
3 - Department of Computer Science, Cornell University
4 - Department of Computer Science, Cornell University
loopp@tc.cornell.edu

LOOPP is a fold recognition program based on the collection of numerous signals, merging them into a single score, and generating atomic coordinates based on an alignment into a homologue template structure. The signals we are using include straightforward sequence alignment, sequence profile, threading, secondary structure and exposed surface area prediction. (Secondary structure and exposed surface prediction program (sable) was developed in the group of our collaborator Professor Jaroslaw Meller). These individual signals are combined locally to create mixed models and globally to provide overall scores. Computations of scores to those that can be done quickly are performed for all proteins in our database and expensive scores (such as Z score calculations) are computed only for those that score highly with the 'cheap' score. Atomic models are then generated using an alignment produced by the scoring scheme and the Modeller program of Andrej Sali. The final atomic structure is evaluated by additional energy scores. The energies used, and the combination of individual scores are determined by a Mathematical Programming algorithm. The final models are processed with MESHI program of Chen Keasar (http://www.cs.bgu.ac.il/~meshi/).

LTB-WARSAW - 397 models for 83 3D/1 TR targets

Automated approach to protein structure prediction with the lattice reduced model and BioShell toolkit suite

D. Gront, S. Kmiecik, M. Kurcinski, D. Latek and A. Kolinski
Warsaw University, Faculty of Chemistry, Pasteura 1, 02-093 Warsaw Poland
dgront@chem.uw.edu.pl

This is a hybrid method that uses threading metaservers and molecular modeling with a reduced representation of the protein conformational space. The results from the servers are subject to 3D-jury (bioinfo.pl) scoring. The top 5-20 templates are selected, depending on the distribution of scores and mutual structural alignment between the templates. The templates are a source of a large number of distance restraints, which are subsequently used in the Replica Exchange sampling optimization with a reduced lattice protein model. We used essentially the same reduced-space CABS[1] modeling tool as the one used by group Kolinski-Bujnicki during the CASP6 experiment, although the force field of the model has been refined (larger database for statistical potentials) and carefully optimized. The whole procedure has been automated with the BioShell[2] package. In template-free targets contacts predicted by servers were applied as week constraints. The best scoring structures are subject to the all-atom rebuilding and a refinement using BBQ[3] and SYBYL methods. For the purpose of the scoring of final models we have tested various MQAP methods and developed a procedure that improves the model by means of all-atom energy minimization. Extensive tests on substantial sets of decoys showed that our selection scheme allows for assessment of the model quality. The top scoring all-atom models were submitted to the CASP7 server.

1. Kolinski A. (2004) Protein modeling and structure prediction with a reduced representation. Acta Biochimica Polonica 51 349-371
2. Gront D. & Kolinski A. (2006) BioShell - a package of tools for structural biology computations. Bioinformatics 22, 621-622
3. Gront D., Kmiecik S. & Kolinski A. (2006) BBQ - Backbone Building from Quadrilaterals. A fast and accurate algorithm for protein backbone reconstruction from alpha carbon coordinates. submitted to J. Comp. Phys.

Luethy - 100 models for 100 DP targets

Consensus distance matrices derived from server predictions used as folding potential

Roland Luethy
roland@luethy.net

Overview

Consensus distance matrices were derived from selected server predictions and then used to guide an *ab initio* folding program. A semi-automated process using the following five major steps was used: 1. models from the servers were ranked and a number of top models were selected; 2. a consensus distance matrix was constructed from these models; 3. the distance matrix was used in the scoring function for *ab initio* folding; 4. full atom models were constructed; 5. the full atom models were ranked and the best model was selected for submission to CASP.

1 Selection of server models:

Only server models that were full length were used. For all full length models the Verify3D [1] scores were computed and from each server the model with the highest score was selected. The selected models were then clustered based on their Cα rmsd into groups with rmsd less than 3 between the structures in a cluster and from each cluster the model with the best Verify3D score was selected. The 10 best models were visually inspected and confirmed. In some cases fewer were used in other cases more models from clusters were reintroduced. The latter was mainly in situations where most models were very similar. The selected models were used in several places later on: to build consensus matrices, as the structural block database for the folding and for the reconstruction of the all atom models.

2 Construction of consensus distance matrices:

Each model selected in step 1 was converted into a Cα distance matrix, the matrices where then averaged and a consensus matrix was constructed with the average distance if the standard deviation was less than 3 and 0 otherwise. If the number of non zero distances was less than 30 times the number of Cα atoms, the standard deviation cutoff was increased to 5. In situations where the number of distances was still less than 30 per Cα atom, the number of models used was reduced by eliminating the models with the lowest Verify3D scores or the models with the largest Cα rmsd from all the other models.

3 Folding:

A simplified structure representation was used for the *ab initio* folding. The simplified models are based on a sequence of internal coordinates: the pseudo torsion angles between four consecutive Cα atoms and pseudo angles between three consecutive Cα atoms. Different structures were generated by randomly selecting blocks from the models selected in step 1 and substituting them into the folding model. To evaluate structures cartesian coordinates for the Cα atoms were reconstructed using constants for all distances and the angles needed to reconstruct the Cα positions. These structures were then evaluated using a potential with a compactness function, a penalty for too close contacts, and a function summing the deviation of the Cα distances from the consensus distance matrix constructed in step 2. Structures were optimized using a simulated annealing protocol to optimize the scoring function and the protocol was run 10 times to generate 10 models. The 10 structures were then submitted to a refinement step in which the pseudo angles were modified by small random changes to further optimize the same folding potential.

4 Final all atom model construction:

First, the coordinates for the Cα and Cβ atoms were reconstructed using constants for the bond lengths. The backbone and sidechain atoms were then reconstructed by selecting the closest five-residue fragment around each residue from the initial models selected in step 1. The all atom structure was then minimized with TINKER[2] using the steepest descent method and a stepwise protocol that kept all C-alpha atoms fixed in the first step and allowed all atoms to move in the last step.

5 Final model selection:

The final models were again ranked by their Verify3D scores and were visually inspected. The model with the best score was submitted.

Conclusion

The approach presented here was based on using the models from the automated prediction servers and was an attempt to capture the best structures or substructures and use them to construct an improved prediction.

1. Luethy R., Bowie J.U., and Eisenberg D. (1992) Assessment of protein models with three-dimensional profiles. Nature. 356(6364): p. 83-5.
2. Ren P. and Ponder J.W. (2002) Consistent treatment of inter- and intramolecular polarization in molecular mechanics calculations. J Comput Chem. 23(16): p. 1497-506.

Ma-OPUS-Quality Assessment - 702 models for 99 3D/100 DP/99 QA targets

Model Quality Assessment Based on a Novel Cα-based Empirical Potential

Yinghao Wu[1], Mingyang Lu[3], Mingzhi Chen[2], Jialin Li[2], and Jianpeng Ma[1,2,3]

[1]*Department of Bioengineering - Rice University, Houston, TX*
[2]*Structural and Computational Biology and Molecular Biophysics - Baylor College of Medicine*
[3]*Verna and Marrs McLean Department of Biochemistry and Molecular Biology - Baylor College of Medicine, Houston, TX*
jpma@bcm.tmc.edu

In order to improve the model quality assessment in structural prediction, we have specifically developed a new C -based empirical potential function. The fundamental motivation is to cope with the fact that many models in prediction are only in the form of C -traces. We specifically tried to avoid using any other information, such as native backbone dihedral angles, since not all cases will have such information readily available.

The total energy consists of six terms:

$$E_{tot} = E_{tertiary} + E_{pairwise} + E_{SR} + E_{HB} + E_{SAS} + E_{3-body}.$$

The first term in the right hand side of the equation, $E_{tertiary}$, is for the tertiary packing energy of two specific tri-peptides with corresponding secondary structure type. The secondary structure types are -helix, -strand or loop, and we use four-letter-code to coarse-grain the amino acid sequence. They are polar (changed and uncharged) and non-polar (small and large) groups. The second energy term, $E_{pairwise}$, is environment-associated distance-dependent pair-wise potential. The environment of a specific C atom is considered as buried or exposed, depended on the number of neighboring C atoms within certain cutoff distance. The third term, E_{SR}, is a short-range energy term. The conformation of each local fragment including five consecutive C atoms is taken into account. This energy term presents the structural preference of certain local fragment. The fourth term, E_{HB}, is an orientation-dependent potential which considers the spatially anisotropic preference of hydrogen bonds. The fifth term, E_{SAS}, is related to the solvent accessible surface of each amino acid. Then, the last term, E_{3-body}, is a three body energy term in order to include the multi-body potential to take into account of effect of all three residues that

make spatial contact in long range. All the statistical distribution is obtained from a structural non-redundant database of non-homologous soluble proteins 1.

Using this new potential function, we were able to recognize 21 out of 25 standard decoy sets 2, which includes four groups: 4state Reduced Decoy sets 3, FISA and FISA-casp3 Decoy sets 4, LATTICE_SSFIT Decoy sets 5, and LMDS Decoy sets 6. To our best knowledge, there is no report in literature of pure C-based potential that reaches this level of performance. Thus, our new potential is a substantial progress in C-coarse-graining level.

In CASP7, we used this scoring function to assess all the server models submitted within 48-hours after the target releasing. The ranking of each model is according to the energy calculated by our C-based potential.

1. Wang G. & Dunbrack, R.L., Jr. (2003) PISCES: a protein sequence culling server. Bioinformatics 19, 1589-91.
2. Tobi D. & Elber R. (2000) Distance-dependent, pair potential for protein folding: results from linear optimization. Proteins 41, 40-6.
3. Park B. & Levitt M. (1996) Energy functions that discriminate X-ray and near native folds from well-constructed decoys. J Mol Biol 258, 367-92.
4. Simons K.T., Kooperberg C., Huang E. & Baker D. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 268, 209-25.
5. Xia Y., Huang E.S., Levitt M. & Samudrala R. (2000) Ab initio construction of protein tertiary structures using a hierarchical approach. J Mol Biol 300, 171-85.
6. Keasar C. & Levitt M. (2003) A novel approach to decoy set generation: designing a physical energy function having local minima with native structure characteristics. J Mol Biol 329, 159-74.

Ma-OPUS-De Novo - 702 models for 99 3D/100 DP/99 QA targets

OPUS: A New *De Novo* Protocol for Determining Overall Protein Topology

Yinghao Wu[1], Mingzhi Chen[2], Jialin Li[2],& Jianpeng Ma[1,2,3]

[1]*Department of Bioengineering - Rice University, Houston, TX*

[2]*Structural and Computational Biology and Molecular Biophysics - Baylor College of Medicine*

[3]*Verna and Marrs McLean Department of Biochemistry and Molecular Biology - Baylor College of Medicine, Houston, TX*

jpma@bcm.tmc.edu

In de novo structure prediction, it is still a monumentally challenging issue to determine the overall topology of relatively large proteins, especially the β-sheet-containing proteins. We developed a suite of novel computational

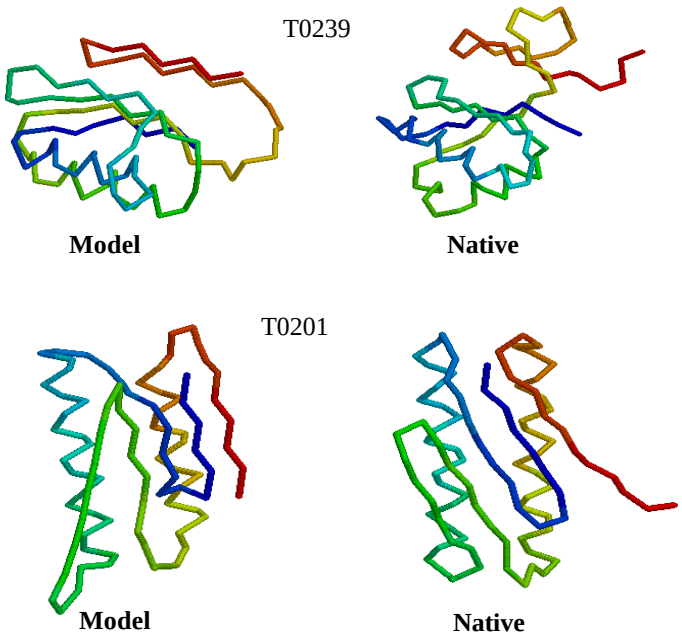


Fig.1 Topology models for two typical hard targets in CASP6. Results are presented together with the native structure.

algorithms to specifically cope with this problem — OPUS, a radically new protocol that determines protein topology by a multi-scale, multi-layer and top-down prediction strategy.

Structural candidates for topological screening were first generated in such a way that the secondary structural elements (SSE) were initially aligned on a predefined lattice space. By that, all the potential topological connectivity can be enumerated. We found that such a novel top-down strategy was very efficient in terms of discrete sampling of protein topology space.

For each topological candidate on lattice, we applied two layers of filter to nail down the native topology. The first, which is also the most important filter, is based on a new β-strand-contact predictor:

Strand Contact Predictor

The core module of our top-down folding method is the -strand-contact predictor. Four types of residues, Val, Ile, Leu, and Phe (VILF), were chosen as characteristic anchors in coarse-graining and determining the -strand-contacts. With VILF-signatures, we derived three topological filters and four statistical scores from the non-redundant protein-database 1. Given a query sequence, all possible -strand-contact sets were first evaluated by filters. Then, the survived sets were ranked by the summations of four scores according to VILF-signatures. We retrospectively tested this -strand-contact predictor on 25 targets up to size of 284aa (taken from the difficult targets in CASP5 and CASP6). There were 16 proteins whose entire set of native -contacts were within top 15 in ranking. The remaining nine targets on average had about 85% of their native -contacts within the top 15 in ranking. The top 15 topological candidates constructed on lattice were sent for further filtering in next step, which was the commonly used 3D-Jury method 2. The -helices, which in most times pack around -sheets, were modeled accordingly based on the beta topology.

In Fig.1, we demonstrate two difficult targets in CASP6 that we got the topology right, while no other group did. We argue that the ability to establish topology, in a de novo sense, for larger proteins is extremely important for pushing structure prediction beyond current level, especially when most of predictions are still assessed by GDT scores 3, which are not sensitive to the correctness of overall topology once the scores are below certain level.

1. Wang G. & Dunbrack R.L., Jr. (2003) PISCES: a protein sequence culling server. Bioinformatics 19, 1589-91.
2. Ginalski K., Elofsson A., Fischer D. & Rychlewski L. (2003) 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 19, 1015-8.
3. Zemla A., Venclovas C., Moult J. & Fidelis K. (1999) Processing and analysis of CASP3 protein structure predictions. Proteins Suppl 3, 22-9.

Ma-OPUS-DOM - 106 models for 99 DP targets

OPUS-DOM: A Novel Method for Domain Boundary Prediction in CASP7

Yinghao Wu[1], Mingzhi Chen[2], Jialin Li[2],& Jianpeng Ma[1,2,3]

[1]*Department of Bioengineering - Rice University, Houston, TX*

[2]*Structural and Computational Biology and Molecular Biophysics - Baylor College of Medicine*

[3]*Verna and Marrs McLean Department of Biochemistry and Molecular Biology - Baylor College of Medicine, Houston, TX*

jpma@bcm.tmc.edu

We have developed a novel computational method for predicting domain boundary. The essence of the method is based on a new coarse-grained *de novo* folding algorithm, called SKELEFOLD, which generates an ensemble of low-resolution structural models by folding the skeletons of protein. Moreover, we have also incorporated three sequence-based filters to give consensus evaluation. By testing the new method on all multi-domain protein targets in CASP6, we obtained overall sensitivity of 75% and specificity of 67%. **The results are substantially better than the published results in literature [1].** Most importantly, our new method predicts the domain boundary in a truly *de novo* sense, i.e., it does *not* rely on any help from sequence homology information. Fig.1 gives a typical example of CASP6 target, T0216, which was one of the most difficult targets in CASP6 new-fold category. The domain boundary was correctly defined by our *de novo* method. Throughout the text please use Times New Roman, 10pt. and single spacing. Please use the following [1-3; 5] citation scheme. Skip one line between paragraphs. No indentations. Please skip one line before the literature block.

T0216 (CASP6)

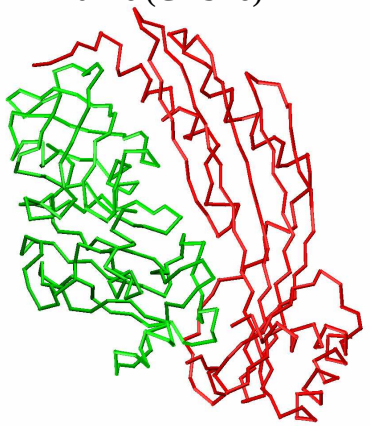


Fig.1 The domain boundary of CASP6 target T0216 is marked on the native structure.

SKELEFOLD Method

SKELEFOLD uses coarse-grained vector representations for secondary structural elements (SSEs), i.e., α-helices, β-strands and loops. A geometry-based scoring function describing packing preference of SSEs in the vector representations was first extracted from a non-redundant protein structure database [2]. Meanwhile, a motif library was

constructed by recording the local internal coordinates of all five-adjacent-secondary-structure fragments from the same database. Then, given the query sequence, a profile-based dynamic programming method was used to select fragment candidates from the library. Guided by the geometry-based scoring function, the initially extended skeleton can be folded into a compact tertiary structural model. Finally, Cα trace was constructed from the vector model.

Domain Boundary Determination

For each query sequence, once SSEs were assigned by PSIPRED [3], we used SKELEFOLD to generate 10,000 compact structural models. The domain boundary for every one of the 10,000 models was analyzed by DOMID software (http://bioinfo1.mbfys.lu.se/Domid/domid.html). Along the sequence, a frequency profile was constructed by recording the occurrence of being identified by DOMID as domain boundary. This profile was then normalized to Z-scores. All the residue positions with Z-score larger than 1.0 were regarded as potential candidates for domain boundary.

For the potential domain boundary candidates generated from SKELEFOLD, three sequence-based filters were applied to give consensus evaluation. 1) Residue entropy index (REI) filter [4] was based on the hypothesis that the domain boundary is conditioned by amino acid residues with a small value of side chain entropy, which correlates with the side chain size. 2) Domain linker index (DLI) filter [5] was derived from the log ratio of the amino acid composition of linker regions to compact domains. 3) We developed a new filter, domain boundary profile library (DBPL). It was to provide the profile information at domain boundary regions from the learning of the same non-redundant structure dataset [2]. The three filters can produce three additional Z-score profiles along the sequence with the more negative value of Z-score the better since the filters are energy-like in nature. Finally, for the potential domain boundary candidates from SKELEFOLD, we would confirm the domain boundary if at least one of the filter indicates a Z-score less than –2.5 within ±15 residues of the candidate boundary position.

We'd like to point out that, although our novel method itself does not have to rely on any homology information and excellent sensitivity and specificity values have been obtained on the CASP6 targets, we used following strategy in CASP7 competition purely for the sake of safety and efficiency on easy targets. We employed a hierarchical screening procedure by using BLAST, PSI-BLAST [6] and threading method FFAS03 [7] to eliminate regions in the query sequence that have obvious domain homologies in known structures, then we applied our new method for the remaining hard regions or the whole hard targets.

1. Kim D.E., Chivian D., Malmstrom L. & Baker D. (2005) Automated prediction of domain boundaries in CASP6 targets using Ginzu and RosettaDOM. Proteins 61 Suppl 7, 193-200.
2. Wang G. & Dunbrack R.L., Jr. (2003) PISCES: a protein sequence culling server. Bioinformatics 19, 1589-91.
3. Jones D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292, 195-202.
4. Galzitskaya O.V. & Melnik B.S. (2003) Prediction of protein domain boundaries from sequence alone. Protein Sci 12, 696-701.
5. Dumontier M., Yao R., Feldman H.J. & Hogue C.W. (2005) Armadillo: domain boundary prediction by amino acid composition. J Mol Biol 350, 1061-73.
6. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W. & Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25, 3389-402.
7. Rychlewski L., Jaroszewski L., Li W. & Godzik A. (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. Protein Sci 9, 232-41.

Meiler - 97 models for 97 PR targets

Contact Prediction Using Artificial Neural Networks

M. Karakaş and J. Meiler

Center for Structural Biology, Vanderbilt University, Nashville, TN

mert.karakas@vanderbilt.edu

Packing of secondary structure elements (SSEs) are shown to be an important aspect in protein folding. Methods predicting contacts between amino acids, in the interfaces of SSEs, provide valuable information helping predict tertiary structure of de novo and hard fold recognition targets[1] . Therefore, a reliable residue-residue contact prediction method based only on sequence information would be able to reduce the conformational search space vastly in de novo fold prediction. We have developed a method where artifical neural networks (ANNs) are trained with data extracted from ~1800 proteins from a non-redundant fold database[2] (less than 25% sequence identity) to differentiate between contacts and non-contacts.

The ANNs require an input of two sequence windows spanning the potentially interacting SSEs, having the two directly contacting amino acids in the center. The length of these sequence windows was chosen to be 9 residues for α-helices and 5 residues for β-strands. In result both SSEs have about the same length of 12 Å for the interaction interface. For each amino acid in these windows, predicted secondary structure (JUFO[3]), position specific scoring

matrices from PSI-BLAST and a property profile are used as input. Five separate ANNs were trained for helix-helix, helix-strand, strand-helix, strand-strand and sheet-sheet interactions.

For an independent set, a fixed threshold has been applied on the probability outcomes from ANNs for decision on whether each residue couple is predicted to be in contact or not. The predictions had an accuracy of 73-79% accuracy (varying depending on the contact type), while 10% of non-contacts were falsely identified as contacts. When looked at predictions with receiver operating characteristic (ROC) curves, the areas under the curves were found to be 78-83%. The contact predictions were also converted to a scoring function and were shown to be successful in differentiating between native-like and non-native structures.

It is expected that high-resolution training of ANNs will increase accuracy of the predictions and result in a further reduction of search space for de novo fold prediction.

1. Grana O., Baker D., MacCallum R.M., Meiler J., Punta M., Rost B., Tress M.L. & Valencia A. (2005) CASP6 assessment of contact prediction. Proteins. 61, Suppl 7:214-224.
2. Wang G. & Dunbrack R.L. (2003) PISCES: a protein sequence culling server. Bioinformatics, 19, 1589-1591.
3. Meiler J. & Baker D. (2003) Coupled prediction of protein secondary structure and tertiary structure. Proc Natl Acad Sci. 100, 21: 12105-12110

Meta-DP - 100 models for 100 DP targets

Meta-DP: Domain Prediction Meta Server

Harpreet Saini[1], Daniel Fischer[2]

[1]-Computer Science and Engineering Department, 201 Bell Hall, University of Buffalo, Buffalo, New York 14260, USA, [2]-Computer Science and Engineering Department, 201 Bell Hall, University of Buffalo, New York 14260, USA

hksaini@cse.buffalo.edu

The Meta-DP domain prediction meta server provides a simple interface to predict domains in a given protein sequence using a number of domain prediction methods. The Meta-DP is a convenient resource because through accessing a single site, users automatically obtain the results of the various domain prediction methods along with a consensus prediction. In addition to the results of individual domain prediction methods, Meta-DP computes and reports consensus prediction using a "majority vote" or a "weighting scheme" in case of a tie. The Meta-DP is currently coupled to eight domain prediction servers and can be extended to include any number of methods. In last CAFASP experiment, Meta-DP was also used to evaluate the performance of thirteen domain prediction methods in the context of CAFASP4-DP. The Meta-DP server is freely available at http://meta-dp.cse.buffalo.

1. Saini H. R. and Fischer D. (2005) Meta-DP: Domain Prediction Meta Server. Bioinformatics 21, 2917-2920.

MetaTasser - 929 models for 100 3D targets

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

S. B. Pandit, H. Zhou, J. Borreguero, S. Lee, H. Chen and J. Skolnick

Center for the Study of Systems Biology, Georgia Institute of Technology

skolnick@gatech.edu

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

1. Ginalski K., Elofsson A., Fischer D. & Rychlewski L. (2003) 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 19, 1015-1018.
2. Zhou H. & Zhou Y. (2004) Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. Proteins 55, 1005-1013.
3. Zhou H. & Zhou Y. (2005) Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins 58, 321-328.
4. Skolnick J., Kihara D. & Zhang Y. (2004) Development and large scale benchmark testing of the PROSPECTOR_3 threading algorithm. Proteins 56, 502-518.
5. Zhang Y., & Skolnick J. (2004) Automated structure prediction of weakly homologous proteins on a genomic scale. Proc. Natl. Acad. Sci. USA. 101, 7594-7599.
6. Zhang Y. & Skolnick J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 33, 2302-2309.
7. Zhang Y., & Skolnick J. (2004) Scoring function for automated assessment of protein structure template quality. Proteins. 57, 702-710.
8. Zhang Y., Kihara D. & Skolnick J. (2002) Local energy landscape flattening: parallel hyperbolic Monte Carlo sampling of protein folding. Proteins. 48, 192-201.
9. Zhang Y., & Skolnick J. (2004) SPICKER: a clustering approach to identify near-native protein folds. J. Comput. Chem. 25, 865-871.

MTUNIC - 483 models for 97 3D targets

Parallel Tempering Monte Carlo based Protein Structure prediction and Refinement

Olav Zimmermann[1], Jan Meinke[1], Sandipan Mohanty[1], Ulrich Hansmann[2]

[1]-Neumann Institute for Computing, Juelich, Germany,
[2]-Neumann Institute for Computing, Juelich, Germany and Michigan Technological University, Houghton, MI
{olav.zimmermann|j.meinke|s.mohanty|u.hansmann}@fz-juelich.de

As a first time participant in CASP our goal was to establish a semiautomatic workflow by combining existing methods for fold recognition with our refinement algorithms and testing certain heuristics for the selection at each step.

Template Selection:

Templates were selected manually from 3D-Jury [1] predictions. Preference was given to high 3D-Jury-scores and agreement between the secondary structure of the template and the predicted secondary structure of the target sequence. For targets which were obviously not CM targets, 3D-Jury predictions from fold recognition servers were preferred. For the second half of the CASP targets we preferred to take the 3-4 templates from different SCOP folds. In few cases where there were no significant 3D-Jury-scores and we suspected that the secondary structure prediction might be wrong we used additionally a fragment based search in the PDB to assess which parts of the PSIpred prediction might be wrong. We did not perform any domain parsing. For a few targets we used templates from the server predictions.

Structure Search:

We searched the fold space [2] employing the CABS program from the Kolinski group [3]. This parallel tempering Monte Carlo program was run using constraints from the respective 3D-Jury templates and secondary structure prediction by PSIpred 2.5 [4]. We used 32 replicas for sequences with less than 200 residues and 64 replicas for proteins with longer sequences. Simulations performed between 15,000 sweeps for long sequences and 100,000 sweeps for short sequences. For each target we used several different constraints settings.

Clustering:

Clustering was performed using hierarchical clustering with HPCM [5] using a fixed difference in RMSD of 2.5 A as clustering radius.

Cluster Selection:

Structure clusters were selected based on cluster averages of CABS energy, and structure similarity (TM-score) to the PDB structure on which the 3D-Jury template was based [6]. Most often we selected those clusters which were in the top 20 for both measures. In ambiguous cases secondary structure content and cluster size was taken into account as well. If too many structures fulfilled the criteria, up to 50 structures were selected manually.

Regularization and Minimization:

Averaged structures from the selected clusters were subject to regularization by SMMP [7]. Regularized structures were ranked according to the total and partial energies of the structures in SMMP, and in particularly ambiguous cases, the consistency of this ranking with a similar ranking based on energy terms of PROFASI [8]. 5 to 10 structures ranked best with this procedure were selected for refinement. For most structures, refinement consisted of a set of constrained simulated annealing runs with SMMP, starting from very high temperatures. Most structures dissolved and reformed into local minima of the potential that were close to the input structures of the refinement procedure. The final structures from different annealing trajectories were once again ranked following a similar procedure as above. In a few cases, local minima

structures obtained from constraint-free parallel tempering runs with PROFASI, starting from random initial states in an all-atom model, were evaluated and ranked based on their partial energies and compactness.

Final Selection:

Final selection and ranking was based on several energy terms, secondary structure content and visual inspection.

1. Ginalski K., Elofsson A., Fischer D., Rychlewski L. (2003) 3D-Jury: a simple approach to improve protein structure predictions, Bioinformatics 19, 1015-1018.
2. Hansmann U.H.E. (1997) Parallel tempering algorithm for conformational studies of biological molecules, Chem. Phys. Lett. 281, 140-150.
3. Kolinski A. and Bojnicki J.M. (2005) Generalized protein structure prediction based on combination of fold-recognition with de novo folding and evaluation of models. Proteins 61 Suppl. 7, 84-90.
4. Jones D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices, J. Mol. Biol. 292, 195-202.
5. Gront D. and Kolinski A. (2005) HCPM—program for hierarchical clustering of protein models, Bioinformatics 21, 3179-3180.
6. Zhang Y., Skolnick J. (2004) Scoring function for automated assessment of protein structure template quality, Proteins 57, 702-710.
7. Eisenmenger F., Hansmann U.H.E., Hayryan S. and Hu C.K. (2006) An enhanced version of SMMP—open-source software package for simulation of proteins, Comp.Phys.Comm. 174, 422-429.
8. Irbaeck A. and Mohanty S. (2006) PROFASI: A Monte Carlo simulation package for protein folding and aggregation, J. Comp. Chem. 27, 1548-1555.

MUMSSP - 19 models for 13 3D targets

Loop refinement and geometry optimization: key steps in protein modeling

M. R. Saberi[1], A. Baratian[1], H. Sadeghian[2]

[1] – *Medicinal Chemistry Department, School of Pharmacy, Mashhad University of Medical Sciences, PO Box: 91775-1365, Mashhad, Iran.*

[2] – *Chemistry Department, Science School, Ferdowsi Univ., Mashhad, Iran*

When modeling proteins, all modelers go through usual procedures *i.e.* searching proper template(s), finding the best alignment(s), predicting the most accurate secondary structure prediction, forecast folding of the protein in super secondary structure and tertiary structure, qualify and assess all gathered data and finally do protein modeling and assessment. They usually go back and improve the models by using different template(s) and alignment(s) and repeat modeling until the best model fulfills them and meet the reality.[1] Lots of sites, servers, computers and software are exploited during a protein modeling project however some modelers develop their own facilities including software and algorithms. The challenge appears when trying to resolve a model for the entire protein including loops.[2] Loops are parts of proteins which fold as they want and can affect the quality and accuracy of a protein.

We report here a deep trial study of loop refinement in improvement of modeled proteins of 13 CASP7 targets. As mentioned, this study utilized usual procedures to fined template(s), alignment, secondary structure prediction, folding prediction, motif prediction, modeling and quality assessment of CASP7 targets. UCLA, NCBI and EBI sites, ExPasy, PDB, FUGUE and PSSM servers and many other bioinformatics web sites and servers as well as software such as MODELLER 8v2, MolMol, ViewerLite, Autodock, Chem3D, Rasmol, *etc.* applied for modeling the targets. What_Check, ERRAT, and verify3D were the methods of protein 3D structure assessment to assess stereochemistry, atom environment and solvent accessibility of models respectively. Trial-error method was the choice until no more improvement was achieved for models. Then models were energy minimized as whole and improper loops separately. Different windows were selected on loops for energy minimization and the windows were shrunk until a few residues remained unrefined .Problematic residues in loops were then selected and minimized in third step until changing the conformation of those residues were not advantageous any more. In the forth step other residues of neighbor segments of the protein in a 3D environment which were not necessarily the neighbor residues in the raw sequence were minimized with the problematic loop residues together in a box using MODELLER's loopmodel class. Finally, the whole proteins were energy minimized by Means of MM+. RMS gradient was decreased in a step wise approach. It was of surprise to see that energy minimization, while improving model's performance in tests dramatically, could damage the structure's performance if excessively applied. This approach could magically refine the problematic loops so that the ERRAT test often raised up to 90-100%. Of course model improvement was tracked during the model refinement applying What_Check, ERRAT, and verify3D methods. 17 CASP7 models submitted by our team looks promising and show high quality compared to the released structures of the targets.

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

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

Nano3D - 316 models for 64 3D targets

Ab initio protein folding

Won Seok Han[1], Min Kyung Lee[1], and Chang No Yoon[2]
[1] *Nanormics, Inc. 10-57 Hawolgokdong, Sungbukku, Seoul, Korea,*
[2]*Korea Institute of Science and Technology, Seoul, Korea*
wshan@nanormics.com,cody@kist.re.kr

Our method for protein structure prediction composed of three parts; "local structure prediction" which determines the structure of 4 consecutive amino acids; "global structure prediction" performed by flexible score (Disorder) and burial score of each amino acid that forms protein; construction of complete protein structure by "domain-domain docking" between 2 domains or fragments (composed of more than 20 amino acids).

In order to predict local structure, we abstracted fragments of 4 consecutive amino acids from PDB, and then constructed local structure database that classified sequence, structure and environment data. When query sequence is given, we pull out the local structure with the highest score from local structure database. We define the local structure as 9 structures; alpha-helix, near- alpha-helix (2), extend (3), and coil (3).

For prediction of location (core or surface) of each amino acid that forms protein, we used flexibility and burial of amino acid. We placed amino acid with high flexible score and low burial score on the surface, while amino acid with low flexible score and high burial score is placed in core.

Since the part with very high flexible score (mostly coil) have various structures possibly, no folding process is carried out for this part. Structure prediction for flexible part would not only be inaccurate but would also have negative effects on the structure prediction for the other parts.

For one protein sequence, when several domains and fragments (composed of more than 20 amino acids) are acquired, instead of one entire structure, completion of the protein structure is done by "domain-domain docking". The initial structure is set so that the cores (where amino acids with low flexible score and high burial score is distributed) of the domains come in contact with each other, then we search for the structure which would have the highest folding score. After domain-domain docking carried out, loop modeling is done.

1. Berman H.M., Westbrrok J., Feng Z., Gilliland G., Bhat T.N., Weissing H., Shindyalov I.N., Bourne P.E. (2000) The Protein Data Bank. Nucleic Acids Res. 28, 235-242.

NanoDesign - 322 models for 82 3D/1 TR targets

Sidechain optimization using NanoDesign

Taesung Moon[1], Minsu Han[1] and Chang No Yoon[2]
[1] *- Nanormics, Inc. 10-57 Hawolgokdong, Sungbukku, Seoul, Korea*
[2] *- Korea Institute of Science and Technology, Cheongryang, Seoul, Korea*
iris@nanormics.com

The tertiary structures for CASP7 targets were generated by Nanormics protein modeling system (NanoModel) which generates protein tertiary structures with fully automated manner. The structures of sidechains generated by NanoModel were then optimized by using Nanormics protein design engine (NanoDesign). In the optimization procedure, the sidechain conformations were taken from Dunbrack backbone-dependent rotamer library[1]. Because the number of rotamers is very large, the Dead-End Elimination (DEE) algorithm[2,3] was applied to reduce the number of them. Rotamer/rotamer pair energies and rotamer/template energies were calculated using AMBER forcefield. A pair of atoms was defined as clashing if their van der Waals energy is greater than 3.0 kcal/mol and all rotamers that clash with the template were excluded. Using the reduced rotamers, the energies of protein were calculated in all possible mutations of considered residues which were generated by exchanging one rotamer for another. The energy terms included in the calculations were van der Waals, electrostatic, and hydrogen bond interactions. A Lennard-Jones 12-6 potential were used for van der Waals interactions and the van der Waals radii of all atoms were scaled by 0.9. A distance dependent dielectric constant was used for electrostatic interactions. Hydrogen bonds were represented by 12-10 potential which is dependant on distance and angle.

1. Dunbrack R.L. Jr & Cohen F.E. (1997) Bayesian statistical analysis of protein sidechain rotamer preferences. Protein Science 6, 1661-1681.
2. Desmet J., De Maeyer M., Hazes B. & Lasters I. (1992) The dead-end elimination theorem and its use in protein side-chain positioning. Nature, 356, 539-542.
3. Goldstein R.F. (1994). Efficient rotamer elimination applied to protein side-chains and related spin glasses. Biophys. J. 66, 1335-1340.

NanoModel - 492 models for 100 3D targets

NanoModel: Protein structure modeling pipeline

Jin Kak Lee[1, 2], Han Su Choi[1] and Chan No Yoon[1,2]
[1] - *Nanormics, Inc. 10-57 Hawolgokdong, Sungbukku, Seoul, Korea*
[2] - *Korea Institute of Science and Technology, Cheongryang, Seoul, Korea*
ljk@nanormics.com

Protein structure modeling pipeline, NanoModel combines the results of sequence alignments and fold recognition alignments to find suitable templates. Model building is carried out using Junctional Fragment Matching (JFM) method, and created models are evaluated by solvent accessibility and residue-residue contact scores.

Template structure identification

To identify a template structure, we used five iterations of PSI-BLAST[1] against protein sequence database. In case that templates are uncertain, we used fold recognition method which searches sequence structure alignment by dynamic programming algorithm, using sequence property and secondary structure.

Protein modeling

Insertion/deletions parts in sequence structure alignment are modeled using similar fragment from Protein Data Bank. Side-chains of conserved residue are fixed and the others are adjusted by side-chain rotamer library. The energy minimization was then performed.

Model evaluation

From multiple sequences alignment we get consensus buried and exposed regions. Residue-residue contact prediction is achieved by neural network. Then, model structures are evaluated by the criteria set with solvent accessibility and residue-residue contact scores.

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

nFOLD - 500 models for 100 3D targets

Fully automated protein fold recognition using a modified version of the nFOLD protocol

L.J. McGuffin[1]
1 – The Bioinformatics and Systems Biology Unit, The BioCentre, The University of Reading, Whiteknights, Reading RG6 6AS, UK
l.j.mcguffin@reading.ac.uk

Tertiary structure predictions were submitted in the automatic server category using a modified version of the original nFOLD protocol. The original version of nFOLD[1,2] aimed to extend mGenTHREADER[2,3] through the incorporation of three additional inputs to the underlying neural network. These extra inputs included; the Secondary Structure Element Alignment (SSEA) score, a model quality assessment score from MODCHECK[4] and a functional site detection score, from a modified version of the MetSite[5] method, which was used to evaluate whether or not the functionally important residues were correctly positioned in the model. The neural network of the original version of nFOLD was trained on model quality scores using the MaxSub[6] method, in an attempt to optimize ranking of models.

The original method worked to some extent in that it showed some improvement over mGenTHREADER in CASP6, on the harder targets. Remarkably, the method also provided one of the best predictions overall for the new fold target T0248 (domain 2)[2]. However, the improvement on hard targets appeared to be offset by the performance on easier targets where no real improvement was shown. It was clear that both the functional site scoring and the training of the method using the MaxSub scores were not optimal, therefore a few improvements have been made to the new version.

The new version of nFOLD essentially maintains the original idea, in that it attempts to select the best models built from mGenTHREADER alignments using a number of different scores. However, the MetSite score has been removed and replaced by two new scores - ProQ-LG and ProQ-MX - obtained from the ProQ[7] method for model quality assessment. In addition the neural network for the new version of nFOLD is trained to rank models based on the TM-scores[8].

1. Bryson K., McGuffin L.J., Marsden R.L., Ward J.J., Sodhi J.S. & Jones D.T. (2005) Protein Structure Prediction Servers at University College London. Nucleic Acids Res. 33, W36-8.
2. Jones D.T., Bryson K., Coleman A., McGuffin L.J., Sadowski M.I., Sodhi J.S. & Ward J.J. (2005) Prediction of novel and analogous folds using fragment assembly and fold recognition. Proteins. 61 (S7), 143-51.

3. McGuffin L.J. & Jones D.T. (2003) Improvement of the GenTHREADER method for genomic fold recognition. Bioinformatics. 19, 874-881.
4. Pettitt C.S., McGuffin L.J. & Jones D.T. (2005) Improving sequenced based fold recognition by use of 3D model quality assessment. Bioinformatics. 21, 3509-3515.
5. Sodhi J.S., Bryson K., McGuffin L.J., Ward J.J., Wernisch L. & Jones D.T. (2004) Predicting metal binding sites in low resolution structural models. J. Mol. Biol. 342, 307-320
6. Siew N., Elofsson A., Rychlewski L. & Fischer D. (2000) MaxSub: an automated measure for the assessment of protein structure prediction quality. Bioinformatics. 16, 776-85.
7. Wallner B. & Elofsson A. (2003) Can correct protein models be identified? Protein Sci. 12, 1073-1086.
8. Zhang Y. & Skolnick J. (2004) Scoring function for automated assessment of protein structure template quality. Proteins, 57, 702-710.

NN_PUT_LAB - 279 models for 94 3D/ 93 DP targets

DomAnS method – the new approach used for predicting domains boundaries in proteins

J. Błażewicz[1, 2], P. Łukasiak[1, 2], M. Miłostan[1], W. Jaś kowski[1]

[1] - Institute of Computing Science, Poznań University of Technology, Piotrowo 2, 60-965 Poznań, Poland and [2] - Institute of Bioorganic Chemistry, Polish Academy of Sciences, Noskowskiego 12/14, 61-704 Poznań, Poland
author1@institution1.edu

The new method of prediction of proteins domains boundaries called DomAnS has been proposed. The DomAnS approach predicts protein domains using a combination of information in the form of templates, fragments patterns and segments. The templates are chains of length from two to twenty amino acids. Each template represents, in the middle of the chain, domain boundary ("cut place"). Taking into consideration all possible combinations of cut places, there are four main types of templates. The first type, domain-domain template (DD), contains the domain boundary from both sides of the cut place. The second type, domain-fragment template (DF), encloses the domain boundary from the left side of the cut place and a fragment from the right side. The fragment-domain template (FD) is a reverse of DF template type. The last type of the template, domain-fragment-domain template (DFD), contains the domains boundaries at the ends of the template and the fragment between these domains boundaries. The fragments patterns are chains of amino acids which are not classified to any protein domain. They have length from one to even several hundred amino acids. The segments are the pieces of protein domain. Each segment contains the templates of start and end of each discontinuous domain.

All these combination of information is stored in database created specially for the DomAnS method. Templates, fragments patterns and segments from this database are derived from four domain classification databases: Dali Domain Dictionary (Holm and Sander), CATH (Orengo et al.), SCOP (Murzin et al.) and Pfam (Sanger Institute). The first three databases contain detailed and comprehensive description of the structural classification and evolutionary relationships among all proteins whose structure is known. These proteins structures can be found in Protein Data Bank (PDB). The Pfam database contains only a collection of multiple sequence alignments and hidden Markov theoretical models covering many protein domains whose structure is not known. These structures can not be found in PDB.

The DomAnS approach first tried to adjust all possible templates of length from eight to twenty amino acids with protein input sequence. After that, fragments patterns are analyzed. The aim of this part of process is to remove all templates with cut place which is on the boundary between domain and fragment. Moreover, any of the fragments patterns can not be aligned to fragment part from analyzed template. At the end of the DomAnS method existence of discontinuous domains are checked by using, stored in the database, the segments.

NN_PUT_LAB - 279 models for 94 3D/ 93 DP targets

3D Judge – meta predictor for 3D protein structure

J. Błażewicz[1, 2], P. Łukasiak[1, 2], M. Antczak[1], M. Miłostan[1], G. Palik[1]

[1] - Institute of Computing Science, Poznań University of Technology, Piotrowo 2, 60-965 Poznań, Poland and [2] - Institute of Bioorganic Chemistry, Polish Academy of Sciences, Noskowskiego 12/14, 61-704 Poznań, Poland
author1@institution1.edu

3D-Judge is a selector meta-predictor. It produces exactly one model as its output having N models (from a set of different N servers) as its inputs. The decision which model should be produced as the output is made based on the following information:

1. The similarity matrix (similarity of pairs of models produced by individual servers).

2. Historical data (models and its evaluations).

3D-Judge uses an artificial neuron network (ANN) in order to choose the best model among models produced by predictors. Each of NxN input neuron of ANN is assigned to one similarity matrix coefficients. ANN has N otput neurons. As historical data (on which ANN was taught) we have used CASP6 publicly available models. We have used GDT (Global Distance Test) as similarity measure between two models. As ANN we have used FANN (Fast Artificial Neural Network Library).

3D-Judge uses the following servers: mGenThreader, GenThreader, nFOLD, FUGUE2, LOOPP, ZHOUSPARKS2, zhousp3, PROTINFO, ESyPred3D.

1. Zemla A. (2003) LGA - a Method for finding 3D similarities in protein structures. Nucleic Acids Research. 31,3370-3374.
2. Nissen S. (2003) Implementation of a fast artificial neural network library (FANN), Department of Computer Science University of Copenhagen, The university report.

Oka - 206 models for 4 3D/100 DP/99 DR targets

Entropy capacity determines protein folding rate

O.V. Galzitskaya and S.O. Garbuzynskiy

Institute of Protein Research, Russian Academy of Sciences

ogalzit@vega.protres.ru

Search and study of the general principles that govern kinetics and thermodynamics of protein folding generate a new insight into the factors controlling this process. Here, based on the known experimental data and using theoretical modeling of protein folding[1], we demonstrate that there exists an optimal relationship between the average conformational entropy and the average energy of contacts per residue, that is an entropy capacity[2], for fast protein folding. Statistical analysis of conformational entropy and number of contacts per residue for 5818 protein structures from four general structural classes[3] (all-α, all-β, α/β, α+β) demonstrates that each class of proteins has its own class-specific average number of contacts (class α/β has the largest number of contacts) and average conformational entropy per residue (class all-α has the largest number of rotatable angles φ, ψ and χ per residue). These class-specific features determine the folding rates: all-α proteins are the fastest folding proteins, then follow all-β and α+β proteins, and finally α/β proteins are the slowest ones. Our result is in agreement with the experimental folding rates for 60 proteins[4]. This suggests that structural and sequence properties are important determinants of protein folding rates.

1. Finkelstein A.V. & Badretdinov A.Ya. (1997) Rate of protein folding near the point of thermodynamic equilibrium between the coil and the most stable chain fold. Fold. Des. 2, 115-121.
2. Galzitskaya O.V., Surin A.K. & Nakamura H. (2000) Optimal region of average side-chain entropy for fast protein folding, Protein Sci. 9, 580-586.
3. Murzin A.G., Brenner S.E., Hubbard T. & Chothia C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures, J. Mol. Biol. 247, 536-540.
4. Ivankov D.N. & Finkelstein A.V. (2004) Prediction of protein folding rates from the amino acid sequence-predicted secondary structure, Proc. Natl Acad. Sci. USA, 101, 8942-8944.

Pan - 586 models for 100 3D/73 FN targets

'Threading' with structural profile

Y. He[2], X.M. Pan[1, 2, *]

[1]*-Department of Biological sciences and biotechnology, Tsinghua University, Beijing, China,* [2]*-National Laboratory of Biomacromolecules Institute of Biophysics, Chinese Academy of Sciences, Beijing, China*

pan-xm@mail.tsinghua.edu.cn

Since homology modeling is still the most effective method for building structures, we have focused on the issues of template searching and alignment as well as evaluation.

In this CASP, our strategy is straightforward. First, to investigate whether the target is a multi-domain protein or whether it has conserved domains, two different methods are used – searching NCBI CDD[1] database by RPS-BLAST and searching PFAM[2] database by HMMPFAM in HMMER[3] package. If there is no significant evidence to confirm the target has conserved domains, then it may be a protein with new fold. Or if the target has multiple domains, it will be split into single domains. Second, the target or domain sequence is searched against NR sequence database by PSI-BLAST[4] in multiple iterations to generate PSSM profile, and then the profile is used in searching against PDBAA sequence database to find available templates in structure library. The HSPs found in PSI-BLAST result are amended according to the S2C[5] database and then sorted in the order of sequence identity and coverage. If there are suitable templates found, they are used in the modeling procedure directly. Or, if not, 'threading' launches. In our '2D-threading' method, structural profile of the target is predicted by our prediction program, and then this profile is searched against a pre-compiled database of structural profiles of representative PDB[6] or SCOP[7] by our alignment program. Suitable templates with highly-

conserved structural information may be filtered out, and the alignments are used for model building by MODELLER[8].

The structural profile is the most important part in the 'threading' method. In order to improve the selectivity of 'threading', much more useful information including PSSM, secondary structure and relative solvent accessibility is taken into consideration. Our previous study indicates that there is a normal distribution of *psi* angles, so we can assign different torsion status for each residue, and this status is also combined into the profile. The structural information for each residue in the target is predicted by our prediction program with the multiple-linear-regression (MLR) algorithm which has been reported previously[9].

Both global and local algorithms are implemented in the alignment routine. We think global algorithm may be better in the domain-domain aligning, since the local algorithm usually falls into a small fragment when there is large diversity in domains. For the targets which have templates detectable by PSI-BLAST but with very low identity, structural profile based 'threading' can improve the alignments between the targets and templates, and make them more reasonable.

The profile contains several types of information, and it is very headachy in the evaluation of the alignments. A simple score system has been applied for the multi-factor evaluation, it can distinguish good from bad, but is difficult to distinguish which is better or worse, so human-intervention is very necessary.

1. Marchler-Bauer A., Anderson J.B., Cherukuri P.F., DeWeese-Scott C., Geer L.Y., Gwadz M., He S., Hurwitz D.I., Jackson J.D., Ke Z., Lanczycki C., Liebert C.A., Liu C., Lu F., Marchler G.H., Mullokandov M., Shoemaker B.A., Simonyan V., Song J.S., Thiessen P.A., Yamashita R.A., Yin J.J., Zhang D. & Bryant S.H. (2005) CDD: a Conserved Domain Database for protein classification. Nucleic Acids Res. 33(Database Issue), D192-6.
2. Finn R.D., Mistry J., Schuster-Bockler B., Griffiths-Jones S., Hollich V., Lassmann T., Moxon S., Marshall M., Khanna A., Durbin R., Eddy S.R., Sonnhammer E.L. & Bateman A. (2006) Pfam: clans, web tools and services. Nucleic Acids Res. 34(Database Issue), D247-51.
3. http://hmmer.wustl.edu/
4. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W. & Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389-3402.
5. http://dunbrack.fccc.edu/Guoli/s2c/index.php
6. Noguchi T., & Akiyama Y. (2003) PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB) in 2003. Nucleic Acids Res. 31(1), 492-3.
7. Andreeva A., Howorth D., Brenner S.E., Hubbard T.J.P., Chothia C. & Murzin A.G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. Nucl. Acid Res. 32(Database Issue), D226-9.
8. Šali A. & Blundell T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol. 234, 779–815.
9. Qin S.B., He Y. & Pan X.M. (2005) Predicting protein secondary structure and solvent accessibility with an improved multiple linear regression method. Proteins. 61(3), 473-80.

Panther - 222 models for 66 3D targets

Alignment and Regularization in Modeling

R.W. Harrison[1]

[1] –*Departments of Biology and Computer Science Georgia State University*
rwh@gsu.edu

Analysis of our results in previous CASP experiments showed several areas where improvement was necessary. These included: sequence alignment, error conditioning and stability in model building algorithms, and the molecular force fields. Improvements in these areas were studied during the interval since CASP-6.

Sequence alignment was done with a novel profile-profile algorithm that combined a correlation measure[2] with a sharpening kernel to enhance the signal to noise ratio in the scoring matrix. Tests done using alignments derived from the FATCAT server[3] showed that this combination of measures was able to handle low identity homologies without explicit gap penalties. The improvement in signal to noise is clearly visible when the cost matrices are displayed as images. Profiles were pre-computed using Psi-blast[1] for the unique chains in the PDB. These profiles were searched with a rapid FASTA-like algorithm which searched for short continuous alignments. These short alignments were logged and full dynamic programming was used to find the final alignments. Alignments were scored with a Z-score that was derived from the score along the alignment vs. all other possible alignments. Z-scores > 2 were indicative of a good alignment, but manual inspection of the cost matrix as an image would occasionally reveal alignments that were meaningful at lower quality (targets 348,363,372). Generally speaking, if any homologies were detected, then many homologies were detected and the difficulty became which homolog to choose.

Any physically realistic molecular mechanics force field must show ill-conditioning because it must have translational and rotational invariance. Therefore an error in a small number of atomic positions can propagate into shifts in position for a large number of atoms. This effect can cause relatively

large and somewhat random distortions in the atomic coordinates when building a homology model. Several different regularization algorithms were developed, tested and applied. The simplest regularization algorithm is to apply harmonic restraints to the coordinates of atoms with approximately known positions from the starting structure. Unfortunately, this approach does not at adapt well to internal collisions and large gaps due to deletions. A more sophisticated regularization algorithm uses unrestrained minimization, in our case conjugate gradients with an inexact step size, and then block superimposes the minimized coordinates on the starting coordinates. This algorithm allows the structure to relax, but is more sensitive than using harmonic restraints. Simulated annealing algorithms in the internal coordinates of a molecule are also better conditioned. A simulated annealing algorithm based on local dominating sets[4] was implemented in AMMP. In this algorithm, a side chain was chosen at random and then the local dominating set surrounding based on residue contacts was derived from the structure. The torsion for this side chain and all the members of its local dominating set were given a random variation followed by block stabilized conjugate gradients for each step of the simulated annealing algorithm. In CASP-7 the initial model was built by combination of an analytic structure builder with a harmonically restrained conjugate gradients energy minimization. Simulated annealing on local dominating sets was used to refine side chain positions, and finally block stabilized conjugate gradients was used to build the final refined model. This procedure is much more stable the pure conjugate gradients, and about 1/3 of the time resulted in small improvements in quality.

It is also necessary to improve the molecular potentials in addition to improving the model building algorithms. A set of 20 very high resolution crystal structures were selected and used as targets for genetic algorithm optimization of the AMMP potential. This resulted in small but measurable improvements in model quality.

1. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W. & Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389-3402.
2. Rychlewski L., Jaroszewski L., Li W. & Godzik A. (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. Protein Sci.;9:232–241..
3. Ye Y.Y. & Godzik A. (2003) Flexible structure alignment by chaining aligned fragment pairs allowing twists.. Bioinformatics 119 suppl. 2. ii246-ii255.
4. Wu W., Du H., Jia X., Li Y. & Huang, S.C-H (2006) Minimum connected dominating sets and maximal independent sets in unit disk graphs, Theoretical Computer Science (TCS), 352(1-3):1-7.

PC2CA - 100 models for 100 QA targets

PC2CA: a pseudocovalent model for protein structures with two centers of interaction per amino acid.

F. Fogolari[1]

1- Dipartimento di Scienze e Tecnologie Biomediche
Universita' di Udine, Piazzale Kolbe, 4, 33100 Udine - Italy
ffogolari@mail.dstb.uniud.it

The quality of models submitted by servers for all CASP7 targets has been evaluated using a discrete empirical forcefield for a reduced protein model termed here PC2CA, because it employs a PseudoCovalent structure[1] with only 2 Centers of interactions per Amino acid.

This model refines a previous empirical potential developed by us[2] by adding specific terms for local backbone and sidechain conformations.

All protein structures in the set top500H[3] have been converted in reduced form. The distribution of pseudobonds, pseudoangle, pseudodihedrals and distances between centers of interactions have been converted into potentials of mean force. A suitable reference distribution has been defined for non-bond interactions which takes into account excluded volume effects and protein finite size.

The correlation between adjacent main chain pseudodihedrals has been converted in an additional energetic term which is able to account for cooperative effects in secondary structure element formation.

Local energy surface exploration is performed in order to increase the robustness of the energy function.

The model and the energy definition proposed have been tested on all the multiple decoys' sets in the Decoys'r'us database. The energetic model is able to recognize, for almost all sets, native-like structures (RMSD less than 2.0 A).

The heterogeneity of the models submitted in CASP7 (e.g. in number of residues modeled, in detail of residue modeling) forced us to adopt additional criteria for ranking the models. For roughly the first half of the targets the ranking was based on the energy per residue, with a weight taking into account global energy.

For the second half of the targets the ranking was strictly based on the global energy.

1. Fogolari F., Cattarinussi S., Esposito G. & Viglino P. (1996) Modeling of polypeptide chains as C alpha chains, C alpha chains with C beta, and C

alpha chains with ellipsoidal lateral chains.Biophys. J. 70, 1183-1197.
2. Berrera M., Molinari H. & Fogolari F. (2003) Amino acid empirical contact energy definition for fold recognition in the space of contact maps. BMC Bioinformatics. 4, 8.
3. Lovell S., Davis I., Arendall W., de Bakker P., Word J., Prisant M., Richardson J. & Richardson D. (2003) Structure validation by calpha geometry:phi, psi and cbeta deviation. Proteins 50:437-450.

Peter-G-Wolynes - 160 models for 32 3D targets

Associative Memory Hamiltonian Protocol for CASP7

Chenghang Zong[1], Joe Hegler[1], Patrick Weinkam[1], Kijeong Kwac[1], Michael Prentiss[1], Peter Wolynes[1]
[1]*-Department of Chemistry and Biochemistry, University of California, San Diego*
czong@ucsd.edu

We initially selected sequences for ab-initio prediction if there were no obvious scaffolds found by the automated comparative modeling servers for threading/comparative modeling. For the selected sequences, we used an Associative Memory Hamiltonian and Water mediated potential (AMW),

with parameters chosen previously by optimization. The optimization aims to produce an energy landscape of the AMW that is as close to an ideal funnel as our reduced model allows without using homology information. The AMH has been optimized separately for all-alpha and alpha-beta proteins. Information from secondary structure prediction was included via a potential that biases the phi-psi angles to the appropriate region of a Ramachandran plot. A sequence dependent hydrogen bond term was used to improve beta sheet formation. Molecular dynamics simulations using this potential were used to select low energy candidate structures. Subsequently, the annealed structures are clustered and a smaller subset of structures was selected for submission using several filters.

The structure prediction protocol we have developed is based on the Associative Memory Hamiltonian (AMH)[1,2,3,4,5,6,7]. Water mediated potentials have been recently developed for alpha proteins [5] and alpha/beta proteins [7].

As a summary, the AMH is intrinsically a coarse-grained model, where each residue is represented C_alpha, C_beta, and O atoms. The Hamiltonian contains three major components: i) sequence-independent polymer physics terms to describe the backbone interactions, ii) sequence-dependent knowledge-based potentials for pairwise residues within short sequence distance, iii) water-mediated potentials for residues in the long sequence distance.

$$H = H_{Backbone} + H_{AM} + H_{Water}$$

The backbone interactions include chain-connectivity, excluded-volume, Ramachandran and chirality potentials. H_{Backbone} = H_{chain} + H_{ev} + H_{rama} + H_{chiral}

The sequence-dependent interactions involve C_alpha-C_alpha, C_alpha-C_beta, and C_beta-C_beta pairs. These interactions are grouped into two proximity classes according to the sequence distance between the interacting residues: short range (3<|i-j|<5) and medium range (5<|i-j|< 8).

A pairwise interaction in the target protein is then associated with the aligned pairwise interactions in memory proteins as follows. Water mediated potentials are designed for interactions between residues with sequence distance: |i-j|>8.

For alpha/beta proteins, beta sheet formation are treated with extra components in Hamiltonian described as H_{Beta} = H_{lc} + H_{hb}. H_{lc} describes loose and weak packing between segments with parallel or antiparallel tendency. H_{hb} describes specific geometry for hydrogen bonds in parallel, antiparallel and hairpin formation.

The above Hamiltonian is optimized for selected sequences respectively for alpha proteins and alpha/beta proteins. Once the energy function is optimized, the minima of the energy function are probed via simulated annealing with molecular dynamics simulations. Our simulated annealing protocol gradually reduces the temperature over a large range as in the tempering of steel in metallurgy. This technique allows for local searches in phase space, hopefully avoiding becoming trapped in a metastable state. We collect all annealed structures and cluster them based on pairwise Q score. The annealed structures are scored by a threading Hamiltonian optimized using an energy landscape strategy [8]. The final selection is primarily based on the threading score, but also incorporated the input from examination of the hydrophobic core, secondary structure packing as well as any available biochemical information.

1. Friedrichs M. S. and Wolynes P. G. (1989) Toward Protein Tertiary Structure Recognition by Means of Associative Memory Hamiltonians, Science 246, 371-373.
2. Hardin C., Eastwood M. P., Luthey-Schulten Z., and Wolynes P. G. (2000) Associative memory Hamiltonians for structure prediction without homology: Alpha-helical proteins, Proc. Natl. Acad. Sci.USA 97, 14235-14240.
3. Hardin C., Eastwood M. P., Prentiss M. C., Luthey-Schulten Z. and Wolynes P. G. (2003) Associative memory Hamiltonians for structure

prediction without homology: α/β proteins, Proc. Natl. Acad. Sci. USA 100, 1679-1684.
4. Eastwood M.P., Hardin C., Luthey-Schulten Z., and Wolynes P.G. (2001) IBM J. Res. & Dev. 45, 475.
5. Papoian G.A., Ulander J., Eastwood M.P., and P.G. Wolynes (2004) From The Cover: Water in protein structure prediction, Proc. Natl. Acad. Sci. USA 101, 3352.
6. Prentiss M.P., Hardin C., Eastwood M.P., Zong C., and Wolynes P.G., Chem J. (2006) Theory Comput. 2, 705.
7. Zong C., Papoian G.A., Ulander J., and Wolynes P.G., Am J. (2006) Chem. Soc. 128, 5168.
8. Koretke K.K., Luthey-Schulten Z., and Wolynes P.G. (1996) Self-consistently optimized statistical mechanical energy functions for sequence structure alignment, Protein Science 5, 1043-1059.

PFP_HAWKINS - 36 models for 36 FN targets

Fully automated GO term prediction with PFP

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

The PFP_HAWKINS automated server for function prediction in CASP7 [http://dragon.bio.purdue.edu/casp_fn/] is a slight variation of the PFP (Protein Function Prediction) server[1] maintained by our group, with output modified to fit CASP7 formatting guidelines. PFP is an automated function prediction server that provides the most probable annotations for a query sequence in each of the three branches of the Gene Ontology (GO). Rather than utilizing precise pattern matching to identify functional motifs in the sequences and structures of these proteins, we designed PFP to increase the coverage of function annotation by lowering resolution of predictions when a detailed function is not predictable. This is ideal for many of the CASP targets.

To annotate a query sequence, PFP extends the functionality of a typical PSI-BLAST search[2] in three distinct ways: first, we extract and score GO annotations based on the frequency of their occurrence in highly similar sequences[3]. The GO is a curated, hierarchical vocabulary describing the function of proteins in three categories: molecular function, biological process, and cellular component[4]. Second, we utilize relatively weak hits produced by a PSI-BLAST query, which are not conventionally used for transfer of function annotation. Weakly similar, lower scoring sequences output by PSI-BLAST are not recognized as orthologs to the query sequence, but often represent proteins sharing a common functional domain. Third, we additionally consider those functions that are strongly associated with the highest scoring annotations as described previously. To score these annotations, we designed a novel data mining tool, the Function Association Matrix (FAM), which quantifies the co-occurrence of GO annotations in proteins whose sequences are included in UniProt. Thus, we can assign function using the FAM that cannot be retrieved directly from PSI-BLAST hits.

The output of the server is the top three highest scoring terms in each of the GO categories, ranked in order of raw score.

1. Hawkins T., Luban S. & Kihara D. (2006) Enhanced automated function prediction using distantly related sequences and contextual association by PFP. Protein Sci. 15, 1550-1556. [http://dragon.bio.purdue.edu/pfp/]
2. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W. & Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389-3402.
3. http://www.ncbi.nih.gov/BLAST/.
4. Martin D.M.A., Barriman M. & Barton G.J. (2004) GOtcha: A new method for prediction of protein function assessed by the annotation of several genomes. BMC Bioinformatics 5, 178.
5. Harris M.A., Clark J., Ireland A. Lomax J., Ashburner M., Foulger R., Eilbeck K., Lewis S., Marshall B., Mungall C., et al. (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 32, D258-D261.

POEM-REFINE - 135 models for 27 3D targets

De novo protein structure prediction by all-atom free-energy refinement with PFF01

S.M. Gopal[1], A. Verma[2], K. Klenin[1] and W. Wenzel[1]
1 – Institute for Nanotechnology, 2 – Institute for Scientific Computing
Wenzel@int.fzk.de

We have recently developed all-atom free-energy forcefields (PFF01/02)[1,2] for de-novo all-atom protein folding. With the combination of efficient optimization methods we are able to predictively fold various proteins from 20-60 amino acids[3,4,5,6] from completely extended structures to 3-4 A RMSD of the native conformation. Even though this approach is much faster than all-

atom molecular dynamics, its computational cost rises steeply with the size of the protein.

To contribute to protein structure prediction we have therefore investigated a low-cost protocol for free-energy refinement that combines a heuristic method for model generation with all-atom scoring in PFF01/02. Conformations generated from different methods are not trivially transferable from one theoretical model to another. In order to obtain a meaningful energy estimate each of conformations must be relaxed in new forcefield to a nearby local minimum. We have pursued a low-cost simulated annealing (50,000 steps, T_{start} =200K T_{final}=2K) for each of the decoy. We cluster the top 50 decoys (lowest in energy) and report the average structure of largest cluster as the prediction. This protocol was tested on the Rosetta decoy set[7] consisting of 32 monomeric proteins. We were able to select the near-native conformations with an average RMSD of 3 Å[8].

Encouraged by theses result we decided to participate in CASP7 for proteins with less than 150 amino acids (because of CPU costs) with a similar protocol comprising three stages:

1) Generation of the decoy set: We have generated the decoy set using the Rosetta++ suite. The method consists of two stages: a) fragment generation using the consensus of secondary structure predictors, b) fragment assembly using ROSETTA[9] algorithm. We generated 5000-10000 decoys for each target (excluding homology).

2) Choosing a decoy-subset for refinement: Since refinement of 10000 decoys exceeded our computational resources in the time-frame of CASP, we clustered the decoys and choose about 1000 decoys from the most populated clusters.

3) Refinement and clustering: We have used the same refinement protocol as described above. The 50 lowest energy decoys were clustered using a hierarchical clustering algorithm. The predictions were chosen from the largest clusters. In absence of a dominant cluster, we chose the predictions from larger clusters and visual inspection.

We were able to generate predictions for 27 targets ranging between 68-146 amino acids. More than half of our targets had no homologs detectable with strong confidence by 3D-JURY. We have quantified our predictions as-high-confidence models (score>=0.4) and low confidence models (score<=0.2), depending on the cluster size.

1. Herges T. and Wenzel W. (2004) An All-atom Force field for Tertiary Structure Prediction of Helical Proteins. Biophysical Journal 87, 3100-3109.
2. Verma A. and Wenzel W. (2006) Stabilization and folding of beta-sheet and alpha-helical proteins in an all-atom free energy model. In preparation.
3. Schug A. and Wenzel W. (2005) Evolutionary Strategies for all-atom folding of the sixty amino acid bacterial ribosomal protein l20. Biophysical Journal 90, 4273-4280.
4. Schug A., Herges T. and Wenzel W. (2004) All-atom folding of the three-helix HIV accessory protein with an adaptive parallel tempering method. Proteins 57, 792-798.
5. Gopal S. M. and Wenzel W. (2006) De-novo folding of the DNA-binding ATF-2 zinc finger motif in an all-atom free energy forcefield. Angew. Chem., Int. Ed. In press.
6. Wenzel W. (2006) Reproducible folding of the trp-zipper. Europhys. Letters In press.
7. Tsai J., Bonneau R., Morozov A.V., Kuhlman B., Rohl C.A. and Baker D. (2003) An improved protein decoy set for testing energy functions for protein structure predtion. Proteins 53,76-87.
8. Verma A. and Wenzel W. (2006) Protein Structure Prediction by All-Atom Free-Energy Refinement. Submitted.
9. Simons K.T., Kooperberg C., Huang E. and Baker D. (1997) Assembly of Protein Tertiary Structures from Fragments with Similar Local Sequences using Simulated Annealing and Bayesian Scoring Functions. J. Mol. Biol. 268, 209-225.

PROFcon-Rost - 77 models for 77RR targets

Prediction of protein residue internal contact through Neural Networks

Marco Punta [1,2] & Burkhard Rost [1,2,3]

[1] CUBIC, Columbia University Bioinformatics Center, Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY 10032, USA; [2] C2B2, Columbia University Center for Computational Biology and Bioinformatics, New York, NY, USA; [3] Northeast Structural Genomics Consortium, New York, NY, USA

mp2215@columbia.edu

Our contact prediction method, PROFcon (Punta and Rost, 2005), combines information from alignments, from one-dimensional predictions, from the region between two contacting residues, and from the average properties of the entire protein chain. The method is based on a simple feed-forward back-propagation neural network (NN). We train the NN on a large number of proteins (748) and validate the method's performance on sets that differ in

protein length, number of aligned homologous sequences, and structural class. While PROFcon performance appears to be rather robust as a function of protein length, it suffers greatly in the absence of a proper number of aligned homologous sequences (sparse evolutionary profiles). The best accuracy is achieved for proteins belonging to the alpha/beta SCOP (Murzin, Brenner et al. 1995) (Andreeva, Howorth et al. 2004) structural class. In the following we give a more detailed description of dataset selection and of the features used as input to the neural network. Note that PROFcon was not retrained after CASP6; hence, the present version is exactly the same used to predict targets at CASP6.

Data sets and cross-validation. The EVA server evaluating structure prediction methods (Koh, Eyrich et al. 2003) maintains a continuously updated subset of sequence-unique PDB chains (no pair of proteins in this set has HSSP-value above 0 (Rost 1999), (Sander and Schneider 1991)). In particular, we use the December EVA release, a set of 3201 protein chains of known structure. From this initial list we remove all non-X-ray structures, all membrane and coiled-coil proteins and proteins with physical chain breaks (Gorodkin, Lund et al. 1999). Then, we divide the X-ray-solved protein list into three sets. For training, we select structures with resolution <=2.0 Ã, for validation (i.e. optimization of all NN parameters), structures with resolution in the interval 2.5-3.0 Ã and for test, structures in the interval 2.0-2.5 Ã. Finally, due to computational limitations, we reduce the test set to include only proteins of length less than 400 aa. Training, validation and test set contain 748, 466 and 633 proteins, respectively.

Definition of contact. Two aa are considered two be in contact if their Cb atoms - Ca for glycines Â– are closer than 8 Ã.

NN architecture overview. We train standard feed-forward NN with back-propagation and momentum term (Rost and Sander 1993). We address the extremely unequal distribution of true (contact) and false (non-contact) by balanced training (Rost and Sander 1993). Since the NN 'sees' the symmetric pairs ij and ji as two different samples, the actual PROFcon output value for the ij pair is obtained as the average over the ij and ji NN output (Pollastri and Baldi 2002). The NN uses 738 input, 100 hidden, and 2 output nodes (contact, non-contact).

Detailed specification of input. The input features encoded into the NN vectors correspond to three different levels of description of the aa pair. The pair is characterized through: 1) local information, 2) connecting segment information, 3) protein information. 1) Local level: ij centered windows and pair-specific features. For each residue pair ij in a protein, the network incorporates information from aa comprised in two windows of size 9 centered around i and j (corresponding to intervals [i-4;i+4] and [j-4;j+4]). Each sequence position within the two windows is characterized by 29 nodes: 20 for the evolutionary profile (i.e. frequency of occurrence of the 20 aa types at that position, as obtained from MSA (Przybylski and Rost 2002), (Rost 1996))), one additional node to account for the N and C terminal residues (Rost and Sander 1993), 4 for the predicted SS (three values per residue for helix-strand-other + one value for prediction reliability), 3 for the predicted SA (two values for buried-exposed + one value for prediction reliability) and, finally, 1 for the conservation weight (Rost 1996). Alignments are obtained through PSI-BLAST (Altschul, Madden et al. 1997) filtering the aligned sequences at 80% sequence identity (i.e. any two sequences in the MSA have <80% sequence identity). SS and SA are predicted by PROFphd (Rost 2004)). Note that we train and test on predicted rather than observed 1D values. As the two windows together account for 18 positions, we need a total of 522 nodes for their description. Two more features are introduced to better characterize the central residues i and j. These are: pair type (hydrophobic-hydrophobic, polar-polar, charged-polar, opposite charges, same charges, aromatic-aromatic, other) (Creighton 1992) (7 nodes) and pair complexity (whether or not the two residues are in a low-complexity region, according to SEG (Wootton and Federhen 1996) (2 nodes). 2) Connecting segment level: central window, length and average properties. The segment's central positions have been shown to be the most informative for contacts (Gorodkin, Lund et al. 1999)). So, we introduce a window of size 5 spanning the interval [int(|i-j|/2)-2; int(|i-j|/2)+2]. Sequence positions within this window are characterized in the same exact way as positions in the ij-centered windows (i.e. 29 nodes each). Further, we use 11 nodes for segment length description, corresponding to sequence separations 6, 7, 8, 9 and to intervals 10-14, 15-19, 20-24, 25-29, 30-39, 40-49, >49 (values chosen by intuition not by optimization). Note that the encoding of segment length was necessary in order to qualitatively reproduce the observed distribution of contact probability versus sequence separation (the shorter the sequence separation, the higher the probability of being in contact) (Fariselli and Casadio 1999). Finally, we add in nodes encoding for segment's average properties: 20 nodes for aa composition, 3 nodes for SS composition and one node for the fraction of aa in the segment in a LCR. Overall, we use 180 nodes for the description of the segment. 3) Protein level: length and average properties. We use 20+3 nodes for the average aa and SS composition of the entire protein, plus 4 nodes to describe the protein length (intervals 1-61, 61-120, 121-240, >241; again, values are chosen by intuition).

1. M. Punta and B. Rost (2005) "PROFcon: novel prediction of long-range contacts" Bioinformatics 21(13): 2960-8.Altschul, S. F., T. L. Madden, et al. (1997).
2. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res 25(17): 3389-402. Andreeva, A., D. Howorth, et al. (2004).
3. "SCOP database in 2004: refinements integrate structure and sequence family data.

4. " Nucleic Acids Res 32 Database issue: D226-9. Creighton, T. (1992). Proteins: Structures and Molecular Properties. Fariselli, P. and R. Casadio (1999).
5. "A neural network based predictor of residue contacts in proteins." Protein Eng 12(1): 15-21. Gorodkin, J., O. Lund, et al. (1999).
6. "Using sequence motifs for enhanced neural network prediction of protein distance constraints." Ismb: 95-105. Koh, I. Y. Y., V. A. Eyrich, et al. (2003).
7. "EVA: evaluation of protein structure prediction servers." Nucleic Acids Research 31(13): 3311-3315. Murzin, A. G., S. E. Brenner, et al. (1995).
8. "SCOP: a structural classification of proteins database for the investigation of sequences and structures." J Mol Biol 247(4): 536-40. Pollastri, G. and P. Baldi (2002).
9. "Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners." Bioinformatics 18(Suppl 1): S62-S70. Przybylski, D. and B. Rost (2002).

ProteinShop - 41models for 9 3D targets

Protein Structure Prediction Using ProteinShop

Nelson Max[1,2] and Silvia Crivelli[2]

[1]Dept. of Computer Science, Univ. of California, Davis, CA 95616, [2]Lawrence Berkeley Laboratory, Berkeley, CA 94720

SNCrivelli@lbl.gov

We describe a novel method to predict the 3D structure of new folds via minimizations of a physics-based energy function. It posits that although fold-recognition servers provide incomplete folding information for targets in the new folds category, this information is valuable for guiding the global optimization process to find the solution.

The method has two phases. Phase I creates a set of initial conformations that have alpha-helices and beta-strands according to secondary structure predictions[1-3]. All alpha-helical proteins are partially folded according to templates obtained from fold recognition meta-servers[4,5]. Proteins that have beta-strands are additionally processed with *BuildBeta*, a ProteinShop[6] function that automatically creates a collection of beta-sheet conformations. The starting conformations are locally minimized and ranked using an all-atom AMBER[7] force field with modified parameters[8], designed to improve its discriminatory ability. Phase II improves these conformations through global minimizations in subspaces of the dihedral angles of amino acids predicted to be coil[9] followed by full-dimensional local optimizations[10].

Method Description: Phase I

Here we construct partially or fully folded initial conformations using secondary structure predictions[1-3]. First, we generate a number of extended conformations featuring alpha-helices and beta-strands according to those predictions. ProteinShop generates the three-dimensional coordinates of an extended protein structure containing alpha-helices and beta-strands using sequence and predicted secondary structure information only. These extended conformations are folded using model templates and ProteinShop, which lets users interactively move beta-strands and alpha-helices relative to each other without breaking the protein structure. ProteinShop performs those motions using inverse kinematics techniques on the flexible coil regions. Second, we obtain the templates from the BioInfoBank metaserver[4], which collects structural models from servers and assesses them using the 3D-Jury consensus[5]. The model templates are those hits with the highest *3D-Jury* scores. Additionally, we may include "welded" model templates, built by combining structural information from two templates. Next we build a set of initial conformation as follows:

Constructing Partially- Folded Structures Using Templates: We use ProteinShop to build an initial partially folded structure for every model template and each extended structure by superimposing the latter to the template. This structure superimposition is performed by aligning each rigid-body portion in the extended structure –i.e., each alpha-helix or beta-strand-- to the corresponding portion on the template. The correspondence between the superimposed rigid-body portions is determined by the alignment generated by the meta-server[4]. Often the model templates provide only partial information due to alignment gaps. The protein fragments that correspond to alignment gaps are left extended. Occasionally, we may create additional initial structures containing alignments of the extended parts that seem likely to us. Next, we use BuildBeta to build fully folded models from each partially-folded model. We call the folded fragments *core regions*.

Constructing Fully-Folded Structures Using BuildBeta: BuildBeta generates a collection of potential beta-sheet conformations using statistical scoring functions derived from both protein-fold topology[11] and sequence matching specificity[12]. BuildBeta operates in two possible modes. 1) *with sequence of amino acids and secondary structure prediction information*: it selects the strands to be zipped together into beta-sheets to form different topologies, and then calls zipping routines to automatically create those structures. If there are alpha helices in the sequence BuildBeta moves them away from the beta-sheet to minimize collisions. 2) *with additional information about core regions*: BuildBeta attempts to align those beta-strands outside the core(s) to the sheets that may be present in the core region(s).

Usually, structures created in this phase present steric overlaps that are resolved after local energy minimizations. To obtain a variety of beta conformations, we

let BuildBeta generate most or all of the possible conformations and then we rank them according to their energy value.

Phase II

This phase improves the initial structures by iteratively performing small-dimensional global minimizations in various subspaces of the space of dihedral angles in the coil regions. The method selects a number of low-energy conformations from the list of initial structures and selects small subsets for improvement by global minimizations. A stochastic global optimization procedure finds the best new positions for the chosen dihedral angles while holding the remaining dihedral angles fixed. The global minimizations are followed by local minimizations in the full-dimensional space. The new full-dimensional local minimizers are then merged with those found previously, and the process repeats until a convergence criterion is met[9].

ProteinShop enables the synergistic integration of human knowledge and computer power. We believe this human-in-the-loop approach is necessary to develop a better understanding of the search mechanism being used and to accelerate time to solution.

1. McGuffin L.J., Bryson K. & Jones D.T. (2000) The PSIPRED protein structure prediction server. Bioinformatics 16, 404-405.
2. Karplus K., Barrett C., and Hughey R. (1998) Hidden Markov Models for Detecting Remote Protein Homologies, Bioinformatics 14(10):846—856.
3. Meiler J., Mueller M., Zeidler A. & Schmaeschke F. (2002) JUFO: Secondary Structure Prediction for Proteins. www.jens-meiler.de.
4. Rychlewski L., Fischer D. & Elofsson A. (2003) LiveBench-6: large-scale automated evaluation of protein structure prediction servers. Proteins. 53 Suppl 6, 542-547.
5. Ginalski K., Elofsson A., Fischer D., & Rychlewski L. (2003) 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics. 19(8),1015-1018.
6. Crivelli S., Kreylos O., Hamann B., Max N. & Bethel W. (2004) ProteinShop: A tool for interactive protein manipulation and steering. Journal of Computer-aided Molecular Design. 18, 271-285.
7. Cornell W.D, Cieplak P., Bayly I., Gould I.R., Merz K.M., Ferguson D.M., Spellmeyer D.C., Fox T., Caldwell J.W., Kollman P.A. (1995) A second generation force field for the simulation of proteins, nucleic acids and organic molecules. J. Am. Chem. Soc. 117:5179-5197.
8. Simmerling C., Strockbine B. & Roitberg A.E. (2002) All-atom structure prediction and folding simulations of a stable protein. J. Am. Chem. Soc. 124, 11258-11259.
9. Crivelli S., Bader B., Byrd R., Eskow E., Lamberti V., Schnabel R., Head-Gordon T. (2002) A physical approach to protein structure prediction. Biophysical Journal, 82:36-49.
10. Jiang L., Byrd R., Eskow E., Schnabel R. (2004) A preconditioned L-BFGS algorithm with application to protein structure prediction. Technical Report, Department of Computer Science, University of Colorado.
11. Ruczinski I., Kooperberg C., Bonneau R. & Baker D. (2002) Distributions of beta sheets in proteins with application to structure prediction. Proteins 48, 85-97.
12. Zhu H. & Braun W. (1999) Sequence specificity, statistical potentials, and three-dimensional structure prediction with self-correcting distance geometry calculations of β-sheet formation in proteins. Protein Science 8, 326-342.

PROTEO - 63 models for 62 3D targets

Protein Folding Simulations through low and high resolution models

Vincenzo Villani, Alfonso Cascone

Dipartimento di Chimica, Università della Basilicata,
Via N. Sauro 85, 85100, Potenza.
villani@unibas.it.

The *Rescaled Protein Folding* (RPF) method has been performed developed by means of an efficient HP *self-avoiding walk* model on a periodic lattice, followed by careful simulations at atomic level.[1,4] On the simplified models the global optimization through Monte Carlo Simulated Annealing calculations is performed. The macromolecule is gradually built, while the temperature is slowly lowered. The hydrophobic and disulphide bonds are taken into account. Then, after the scale change and the solvation, the structures are refined through local optimization and molecular dynamics. Lastly, the matching between the simulated and the experimental structure is performed. The method was applied to a number of protein, as BPTI (*Bovine Pancreatic Trypsine Inhibitor*, 57 residues), Ribonuclease (124 residues) and Tyrosin-Kinase domains (about 160 residues). The globular-micellar structure and the gyration radius are obtained. The predictive comparison techniques based on neural networks [5] will be considered to take into account the disulphide bridges and secondary structure patterns.

1. Dill, K. A., Bromberg S. (2002) Molecular Driving Forces, Garland, New York.
2. Chan H. S., Dill K. A. Physics Today (1993), February, 24-32.
3. Villani V., Cascone A. (2004) Recent Res. Devel. Polym. Sci. 8, 21-50.

4. Villani V., Cascone, A. (2005) Recent. Res. Devel. Macromol. 8, 1-24.
5. Pollastri G., Przybylski D., Rost B., Baldi P. (2002) Proteins 47, 228-235.

Pushchino - 4 models for 4 3D targets

Combining sequence alignment tools with threading approach to improve the quality of protein structure prediction

M.Yu. Lobanov[1], N.S. Bogatyreva[1], D.N. Ivankov[1], S.O. Garbuzynskiy[1], O.V. Galzitskaya[1], I.I. Litvinov[2], M.A. Roytberg[2], A.V. Finkelstein[1]

[1]*Institute of Protein Research RAS*

[2]*Institute of Mathematical Problems in Biology RAS*

afinkel@vega.protres.ru

At the first step we launched 10 iterations of standard PSI-BLAST[1] search. From obtained list of proteins, up to e-value of 1000, we selected 120 sequences with the lowest e-values, obtained anywhere during the 10 iterations.

To divide the target by domains we used our program[2] (see abstract of group "Oka") and alignments obtained by PSI-BLAST[1].

Final alignment and selection was done by our home-made program SCF_THREADER[3] with scoring function that takes into account the following factors:

(1) similarity of sequences calculated by similarity matrices GON250 and BLOSUM62;
(2) similarity of secondary structures[4], where target secondary structure was predicted by PSIPRED[5] and template secondary structure was calculated using DSSP[6];
(3) specific energy of short aligned regions that takes into account the type of target amino acids and the conformation of template in this short region;
(4) specific energy of each aligned amino acid pair that takes into account the template residues interacting with the template residue of considered aligned pair;
(5) specific energy of unaligned regions that depends on the conformations of the ends of these unaligned regions and the type of secondary structure of neighbor aligned regions.

1. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25 (17), 3389-3402.
2. Galzitskaya O.V., Melnik B.S. (2003) Prediction of protein domain boundaries from sequence alone. Protein Sci. 12 (4), 696-701.
3. Rykunov D.S., Lobanov M.Y., Finkelstein A.V. (2000) Search for the most stable folds of protein chains: III improvement in fold recognition by averaging over homologous sequences and 3D structures. Proteins 40 (3), 494-501.
4. Litvinov I.I., Lobanov M., Yu., Mironov A.A., Finkelstein A.V. (2006) Information about the protein secondary structure improves quality of an alignment of protein sequences. Mol. Biol. (Moscow) 40 (3), 533-540.
5. Jones D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292 (2), 195-202.
6. Cabsch W., Sander C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 22 (12), 2577-637.

QA-ModFOLD - 100 models for 100 QA targets

ModFOLD: a consensus of model quality assessment programs using an artificial neural network

L.J. McGuffin[1]

1 – The Bioinformatics and Systems Biology Unit, The BioCentre, The University of Reading, Whiteknights, Reading RG6 6AS, UK

l.j.mcguffin@reading.ac.uk

Predictions in the model quality assessment (QMODE 1) category were generated using the newly developed ModFOLD method. The method, which is based on the nFOLD protocol[1], combines the output from a number of model quality assessment programs (MQAPs) using an artificial neural network.

The output scores from MODCHECK[2], ProQ-LG[3], ProQ-MX[3] and ModSSEA are used as inputs to a feed forward back propagation network. The neural network is trained to discriminate between models based on the TM-score[4].

ModSSEA is a new model quality assessment program based on secondary structure element alignments (SSEA). The ModSSEA score is essentially the same as the SSEA score used in the nFOLD protocol, however the PSIPRED[5] predicted secondary structure of the target is aligned against the DSSP[6] assigned secondary structure of the model. ModSSEA was found to be an effective model quality assessment program in its own right, however further accuracy could be gained by using the consensus approach of ModFOLD.

Although the quality assessment category is a manual prediction category the ModFOLD predictions were carried out entirely automatically for all targets. A web server[7] has been implemented for the ModFOLD method, which accepts

gzipped tar files of models and returns predictions in the QA (QMODE1) format via email.

The ModFOLD method is a true MQAP – one model can be assessed at a time and the output score is based on that model alone. The scores and rankings do not depend on the clustering of many different models relating to the same sequence. Thus, each quality score is predicted for each model independently of any other model.

1. Jones D.T., Bryson K., Coleman A., McGuffin L.J., Sadowski M.I., Sodhi J.S. & Ward J.J. (2005) Prediction of novel and analogous folds using fragment assembly and fold recognition. Proteins. 61 (S7), 143-51.
2. Pettitt C.S., McGuffin L.J. & Jones D.T. (2005) Improving sequenced based fold recognition by use of 3D model quality assessment. Bioinformatics. 21, 3509-3515.
3. Wallner B. & Elofsson A. (2003) Can correct protein models be identified? Protein Sci. 12, 1073-1086.
4. Zhang Y. & Skolnick J. (2004) Scoring function for automated assessment of protein structure template quality. Proteins. 57, 702-710.
5. Jones D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol. 292, 195-202.
6. Kabsch W. & Sander C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 22, 2577-637.
7. http://www.biocentre.rdg.ac.uk/bioinformatics/ModFOLD

RAGHAVA-GPS-mango - 285 models for 95 FN targets

MANGO: prediction of Genome Ontology (GO) class of a protein from its amino acid and dipeptide composition using nearest neighbor approach

G. P. S. Raghava
Institute of Microbial Technology
Sector-39A, Chandigarh, INDIA
raghava@imtech.res.in

One of the major challenges in era of genomics is to predict the function of proteins. As number of proteins whose sequence is known is growing with exponential rate due to advancement in DNA sequence techniques. This has pose a major challenge to the boinformatician to develop strategy to predict the function of protein. Fortunately, function of a large number proteins have been deduced using experimental techniques, one may obtained the information about manually annotated proteins from SWISSPROT database. Recently initiatives were taken to provide the uniform definition of class of protein. Genome ontology is one of the major source of information from where one can obtained the information of class of protein. In GO database the annotation of proteins are at three level i) Biological functions; ii) Biological Process and iii) cell. However, a large number of method already developed in past to predict the class of proteins are limited to predict few classes of proteins. In this study we create the dataset of proteins for each class of GO. These proteins were obtained from UNIPROT database where function of these proteins is manually annotated as per GO classification. For each class of GO we create the average composition of proteins belongs to that class. Lets a given GO class have 200 proteins than we compute overall composition of each of 20 the natural residues. This residue composition represents the class. In order to predict the functional class of a query sequence (CASP6 targets), first composition of query sequence is calculated then we compute the Euclidian distance between composition of query sequence and each class of GO. The class having minimum Euclidian distance were assigned as class of query proteins.

It has been shown in past that dipeptide composition have more information than simple composition because order of neighbor is also considered. Thus we implement our approach using dipeptide composition, where dipeptide composition of proteins were used to calculate Euclidian distance between query protein and GO class of proteins instead of residue composition. We also compute the overall difference (residue composition and dipeptide composition) in query and GO class of proteins. In summary we used composition, dipeptide composition and comination of both for predictiog GO class of target proteins.

RAPTOR-ACE - 500 models for 100 3D targets

An integer linear programming based consensus fold recognition method

Libo Yu[1,4], Dongbo Bu[1,2], Shuaicheng Li[1], Xin Gao[1], Jinbo Xu[3,1] and Ming Li[1]

[1]-David R. Cheriton School of Computer Science University of Waterloo Waterloo Ontario, Canada N2L 3G1, [2]– Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, [3]–Toyota Technological Institute at Chicage, Chicage, IL 60637, [4]– Bioinformatics Solutions Inc. ON, Canada. lyu@bioinfor.com, {dbu,scli,x4gao,mli}@cs.uwaterloo.ca,j3xu@tti-c.org

Protein structure prediction is one of the fundamental problems in the bioinformatics. Frequently, the consensus prediction methods outperform others in recent CASPs by combining the strength of individual prediction methods. However, the consensus prediction methods suffer from the "sever correlation" drawback, that is, some servers may be correlated, and the simple "majority vote" rule fails to select the native structure from a template database.

In this paper, a novel consensus method is proposed to reduce the negative effects caused by the correlation among prediction servers. Briefly, the correlation occurs from the fact that some servers tend to generate similar results since they adopt similar techniques, including sequence alignment tools, secondary structure prediction methods, and scoring functions, etc. Suppose, behind the explicit prediction servers, there are some independent hidden servers, each representing a common feature shared by a set of prediction severs; and for a candidate structure, each explicit prediction server assigns it a score based on the scores given by these hidden servers. Therefore, identifying the hidden independent servers is essential to reduce the negative effects of server correlation, and subsequently to design a more accurate scoring function to select a better model.

In our method, we first employ the maximum likelihood technique to estimate the server correlation; then adopt the factor analysis technique to uncover the hidden servers; and finally design an integer linear programming method to derive the optimal weight for each hidden server. Details of each step are described in the following subsections.

1. Maximum Likelihood Estimation of Server Correlation

Let $s_i\ (1 \le i \le n)$ denote a prediction server, $h_j\ (1 \le j \le m)$ denote a hidden server, and $M = \{M_k \mid 1 \le k \le |M|\}$ denote the model database. For a target $T_l\ (1 \le l \le |T|)$, each server s_i yields a set of models $M_{i,l,q}$ $(q = 1..n_{i,l})$ (here, $n_{i,l}$ is the number of models produced by server s_i for target T_l.) as candidate structures.

Since some servers adopt similar alignment techniques and scoring functions, they always produce similar results. Let $p_{i,j}$ denote the probability that for a target, s_i returns a model similar to server s_j. Here, two models are defined to be similar if the distance between them is above a threshold, say, *GDT score* is greater than 0.5. Under a reasonable assumption that that targets $T_l\ (1 \le l \le |T|)$ are mutually independent, the likelihood that servers $s_i\ (1 \le i \le n)$ generate predictions $M_{i,l,q}\ (q = 1..n_{i,l})$ is

$$L(p_{i,j}) = \prod_{l=1}^{|T|} \binom{n_{i,l}}{ov(i,j,l)} p_{i,j}^{\,ov(i,j,l)} (1 - p_{i,j})^{n_{i,l} - ov(i,j,l)}$$

, where $ov(i,j,l)$ is the number of $M_{i,l,q}$ that same to a model returned by s_j for a given target T_l.

Therefore, the maximum likelihood estimation of $p_{i,j}$ can be calculated as follows: $p_{i,j} = \dfrac{\sum_{l=1}^{|T|} ov(i,j,l)}{\sum_{l=1}^{|T|} n_{i,l}}$

In the rest of this paper, we use P to denote the matrix $P = (p_{i,j})_{n \times n}$.

2. Uncovering the Hidden Servers

Typically, a prediction server measures each model M_k in a template database M, assigns it with a score based on a scoring function, and then reports the top ones as candidate structures. For a given target T_l, let $S_{i,k}$ and $H_{j,k}$ be the probability that model M_k is chosen as one of a prediction results by server s_i and h_j, respectively. Since the hidden servers h_j are mutually independent, it is reasonable to assume that $S_{i,k}$ is a linear combination of $H_{j,k}\ (1 \le j \le m)$, that is,

$$\vec{S}_i = \lambda_{i,1}\vec{H}_1 + \lambda_{i,2}\vec{H}_2 + \ldots + \lambda_{i,m}\vec{H}_m, \sum_j \lambda_{i,j} = 1$$

,where $\vec{S}_i =< S_{i,1}, S_{i,2}, \ldots, S_{i,|M|} >$ and $\vec{H}_j =< H_{j,1}, H_{j,2}, \ldots, H_{j,|M|} >$. Here, a higher coefficient $\lambda_{i,j}$ means that server s_i tends to adopt models reported by h_j than others hidden servers.

From the correlation matrix $P = (p_{i,j})_{n\times n}$, factor analysis technique is employed to derive $\lambda_{i,j}$ and $\vec{H}_j$, that is, $\vec{H}_j$ can be represented to be a linear combination of $S_i (1 \le \vec{i} \le n)$ as follows:

$$\vec{H}_j = \omega_{j,1}\vec{S}_1 + \omega_{j,2}\vec{S}_2 + \ldots + \omega_{j,n}\vec{S}_n$$

, where $< \omega_{j,1}, \omega_{j,2}, \ldots, \omega_{j,n} >$ is an eigenvector of matrix $P^T * P$.

3. ILP Model to Weigh Hidden Servers

Having derived the hidden servers $h_j (1 \le j \le m)$, we can design a new prediction server S', the optimal linear combination of the hidden servers. S' assigns each model with a score as follows:

$$\vec{S}' = \lambda'_1\vec{H}_1 + \lambda'_2\vec{H}_2 + \ldots + \lambda'_m\vec{H}_m \quad \ldots\ldots\ldots(*)$$

To determine a reasonable setting of coefficient λ'_j, a training process is conducted on a training dataset $D = \{< T_l, M_l^+, M_l^- >, 1 \le l \le |T|\}$, where T_l is a target, $M_l^+ \in M$ denotes its native models, and $M_l^- \in M$ the set of false models. The training process attempts to maximize the gap between scores of the native models and false models. In more details, for each target T_l in the training dataset, both its native models and false models will be assigned a score by S'; and a reasonable setting of coefficient should assign a native model a score higher than that for the false ones. The larger the gap between scores, the more robust the prediction ability is. In practice, ``soft margin" idea is adopted to take outliers into account; that is, we try to maximize the gap while allowing errors on some samples. Formally, the learning techniques can be formulated into an integer linear programming problem as follows:

$$\max \sum_{l=1}^{|T|} z_l$$

$$s.t. \sum_{j=1}^{m} \lambda_j h_{j,p} - \sum_{j=1}^{m} \lambda_j h_{j,q} \ge x_{p,q} - 1 + \delta, \ \textit{for each } T_l \in T, M_p \in M_l^+, M_q \in M_l^-$$

$$\sum_q x_{p,q} + y_{p,l} \le |M_l^-|, 1 \le l \le m$$

$$\sum_p (1 - y_{p,l}) \ge z_l, 1 \le l \le m$$

$$\sum_{j=1}^{m} \lambda_j = 1$$

$$\lambda_j \ge 0, x_{p,q} = 0/1, y_{p,l} = 0/1, z_l = 0/1.$$

Here, $x_{p,q}$ and $y_{p,l}$ are 0/1 integer variables. The first restriction force $x_{p,q}$ to be 1 if the score of M_p is greater than that of M_q, i.e., $\sum_{j=1}^{m} \lambda_j h_{j,p} - \sum_{j=1}^{m} \lambda_j h_{j,q} \ge \delta$ (here, the constant δ represents the lower bound of gap.). The second and the third restrictions will set z_l if there exists at least a native model M_p that has score greater than all the false models in M_l^-. The objective function aims to maximize the number of targets that are correctly classified.

To predict models for a given target T_l, $S'(M_k)$ is calculated using formula (*), all the models in database are sorted according to $S'(M_k)$, and the top ones will be selected as the consensus results.

RAPTORESS - 500 models for 100 3D targets

An Atom-level Refinement Approach for

Protein Structure Prediction

Xin Gao[1], Ming Li[1] and Jinbo Xu[2]
[1] - *David R. Cheriton School of Computer Science*
University of Waterloo
Ontario, Canada, N2L 3G1
[2] – *Toyota Technological Institute at Chicago*
Illinois, US, 60637
x4gao@cs.uwaterloo.ca

The biennial CASP[1,2,3] experiments have provided an in-depth and objective assessment of the performance of computational protein structure prediction methods. CASPs have greatly accelerated the development of both human expert and automated prediction methods. By analyzing the results published by CASP6[4,5,6,7], we observed that top ranking automated servers can generate reasonably good predictions or at least good regions for most targets; for example, the server RAPTOR[8,9,10] generated models for several targets in CASP6 which had a region with accurately predicted α-helices and a region with accurately predicted β-sheets. However, the whole models were ranked lowly because the relative orientations of these two regions were far away from the native structures. Furthermore, RAPTOR also lost marks on unaligned regions of models because the quality of threading based methods closely depends on the alignments. Thus, a refinement method is urgently needed to improve the accuracy for such models.

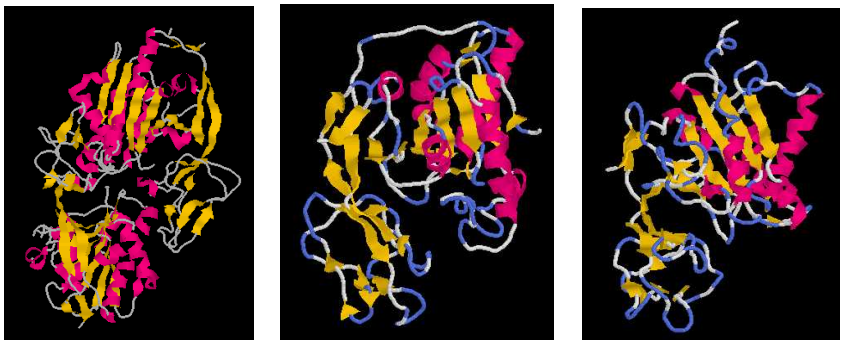
RAPTORESS, our preliminary experiment on an atom-level refinement approach, aims to adjust reasonably good models, which have the whole backbone not too far away from native structures or have some well predicted regions, to get final high-resolution models.

The first refinement stage of RAPTORESS is making all the input models to be protein-like. Some models with well predicted regions rank lowly because they are not protein-like due to the long poorly predicted regions. This step can eliminate the effect of these bad regions. RAPTORESS examines the unaligned regions of models, and uses new regions generated by a comparative modeling based approach to replace those bad regions. The comparative modeling approach searches the region database for a region with the highest score. Then, the bad region in the model is replaced by this good one by translating and rotating the parts at the ends of this region to connect them together. The translation and rotation is directed by an atom-level energy function. The possible rotation angle space is discretized by $10^o \times 10^o \times 10^o$. The conformations with the first two lowest energy scores for each input model are considered to be candidates of final prediction.

The second refinement stage is to adjust all candidate models step by step. This is done on an atom-level on-lattice model. For each iteration, we allow atoms to move within a certain distance. We formulate an integer linear programming (ILP) formulation to restrict the biological and statistical features of models to satisfy constraints of proteins. The conformation determined by ILP is selected to replace the original model in the candidate database. After each iteration, bad models are discarded if the energy is higher than a certain threshold. Then we iteratively repeat this step to explore larger conformational space. The procedure stops either if there is only one model left with energy lower than the threshold, or if we have repeated the step for ten iterations.

After constructing the model set, the models with the first five lowest energy are selected to be the final predictions.

According to preliminary assessment by different websites, such as CAFASP5 assessment, Zhang assessment, and Robetta assessment, RAPTORESS did well in CASP7, especially on some targets, RAPTORESS ranked number one among all the automated servers on the prediction of whole targets or domains. Following is T0289, on which RAPTOESS is ranked number one on whole target by Zhang assessment, and ranked number one on domain two by Robetta assessment. Fig1(a) is the native structure of T0289, which has the PDB code 2GU2. Fig1(b) is the top one model generated by RAPTORESS. Fig1(c) is the top one model generated by RAPTOR, which is an input prediction of the RAPTORESS on this target. For this target, the whole structure is refined to be 3% better on GDT score. The first domain is refined to be 4% better, while the second domain is refined to be 5% better.



(a) (b) (c)

Fig 1 T0289 native structure, Top1 model by RAPTORESS, and Top1 model by RAPTOR

To sum up, RAPTORESS can refine reasonably good models or models with well predicted regions to be more accurate models. The future work will be on more accurate and vehement refinement methods.

1. Moult J., Hubbard T., Fidelis K., & Pedersen J. (1999) Critical assessment of methods on protein structure prediction (CASP) – round III. Proteins: Structure, Function and Genetics, 37(S3), 2-6.
2. Moult J., Fidelis K., Zemla A., & Hubbard T. (2001) Critical assessment of methods on protein structure prediction (CASP) – round IV. Proteins: Structure, Function and Genetics, 45(S5), 2-7.
3. Moult J., Fidelis K., Zemla A., & Hubbard T. (2003) Critical assessment of methods on protein structure prediction (CASP) – round V. Proteins: Structure, Function and Genetics, 53(S6), 334-339.
4. Moult J., Fidelis K., Rost B., Hubbard T., & Tramontano A. (2005) Critical assessment of methods on protein structure prediction (CASP) – round VI. Proteins: Structure, Function and Genetics, 61(S7), 3-7.
5. Tress M., Ezkurdia I., Grana O., Lopez G., & Valencia A. (2005) Assessment of predictions submitted for the CASP6 comparative modeling category. Proteins: Structure, Function and Genetics, 61(S7), 27-45.
6. Wang G., Jin Y., & Dunbrack Jr. L. (2005) Assessment of fold recognition predictions in CASP6. Proteins: Structure, Function and Genetics, 61(S7), 46-66.
7. Vincent J., Tai C., Sathyanarayana B., & Lee B. (2005) Assessment of CASP6 predictions for new and nearly new fold targets. Proteins: Structure, Function and Genetics, 61(S7), 67-83.
8. Xu J., & Li M. (2003) RAPTOR: optimal protein threading by linear programming. Journal of Bioinformatics and Computational Biology. 1(1), 95-117.
9. Xu J., Xu Y., & Li M. (2004) Protein threading by linear programming: theoretical analysis and computational results. Journal of Computational Optimization. 8(4), 403-418.
10. Xu J., Li M. (2003) Assessment of RAPTORs linear programming approach in CAFASP3. Proteins: Structure, Function, and Genetics. 53(S6), 579-584.

ROBETTA - 495 models for 99 3D targets

Robetta *De Novo* and Homology Modeling in CASP7

D. Chivian[1], D. E. Kim[2], B. Qian[2], R. Das[2] and D. Baker[2]

[1] – Lawrence Berkeley National Laboratory, Berkeley, CA

[2] – University of Washington, Seattle, WA

DCChivian@lbl.gov

The Robetta server[1,2,3] (http://robetta.org) combines the Rosetta homology modeling[4] and *de novo*[5] tertiary structure prediction protocols with the Ginzu[1,6] homolog identification and domain parsing protocol to provide predictions for the full length of each target. As a new approach, we modified the Robetta homology modeling protocol from that used in previous CASPs to combine a consensus score with energetic selection from a model ensemble[4]. Our model ensembles are parametrically generated for up to 5 parents by the K*Sync[4] alignment method for the template regions and with Rosetta loop modeling[7] for unaligned regions. Additionally, we modified the Robetta *de novo* protocol to allow for longer trajectories in the generation of each decoy. Blind benchmarking of servers is crucial as it allows us to measure the abilities of automated prediction, vital for the purpose of large-scale prediction efforts.

Robetta homology modeling protocol

Robetta uses up to 5 of the highest confidence detections from BLAST/PSI-BLAST[8] or 3DJury-A1[9] to select the parent for homology modeling. Important to note is that **Robetta does not use the alignment from the detection method** except to determine the domain(s) of the parent to model against. Rather it parametrically generates its own alignment ensemble using the K*Sync alignment method[4] by varying the sequence profile comparison method, the source of the secondary structure prediction, the stringency of the sequence profile, the stringency of the StrAD-Stack[4] multiple structural alignment used to define obligate elements, and the weights on the terms in the dynamic programming scoring function. The alignment ensemble is turned into a decoy ensemble by placing the sequence of the query onto the backbone of the parent based on the alignment. Unaligned loop regions are assembled from fragments and optimized to fit the aligned template structure[7,10]. The template region is kept fixed, and models are selected from the ensemble using a combination of the Rosetta energy function with a consensus score derived from the alignment ensemble[4].

Robetta *de novo* protocol

Robetta *de novo* modeling generates 4000 query decoys and 2000 decoys each for up to 2 homologous sequences (filtered down to 2000, 1000, 1000 to ameliorate known pathologies such as low contact-order structures) using the Rosetta fragment-assembly methodology[5]. Earlier versions of the protocol

generated a greater number of decoys, but had shorter trajectories. We took advantage of the increased resources generously made available to us by the NCSA for the experiment to investigate whether longer trajectories would produce superior decoys. The filtered ensemble is structurally clustered, and the top 5 cluster centers by population are returned in order as the final predictions. Side-chains are added using a backbone-dependent rotamer library[11] with a Monte Carlo conformational search procedure[12].

1. Chivian D., Kim D.E., Malmstrom L., Bradley P., Robertson T., Murphy P., Strauss C.E., Bonneau R., Rohl C.A., & Baker D. (2003) Automated prediction of CASP-5 structures using the Robetta server. Proteins 53, 524-533.
2. Kim D.E., Chivian D., & Baker D. (2004) Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res 32, W526-W531.
3. Chivian D., Kim D.E., Malmstrom L., Schonbrun J., Rohl C.A., & Baker D. (2005) Prediction of CASP6 structures using automated Robetta protocols. Proteins 61, 157-166.
4. Chivian D. & Baker D. (2006) Homology modeling using parametric alignment ensemble generation with consensus and energy-based model selection. Nucleic Acids Res. Sep 13 [Epub].
5. Bonneau R., Strauss C.E., Rohl C.A., Chivian D., Bradley P., Malmstrom L., Robertson T., & Baker D. (2002) De novo prediction of three-dimensional structures for major protein families. J Mol Biol 322, 65-78.
6. Kim D.E., Chivian D., Malmstrom L., & Baker D. (2005) Automated prediction of domain boundaries in CASP6 targets using Ginzu and RosettaDOM. Proteins 61, 193-200.
7. Rohl C.A., Strauss C.E., Chivian D., & Baker D. (2004) Modeling structurally variable regions in homologous proteins with rosetta. Proteins 55, 656-677.
8. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., & Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25, 3389-3402.
9. Ginalski K., Elofsson A., Fischer D., & Rychlewski L.. (2003) 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 19, 1015-1018.
10. Canutescu A.A., & Dunbrack R.L. Jr. (2003) Cyclic coordinate descent: A robotics algorithm for protein loop closure. Protein Sci 12 963-972.
11. Dunbrack R.L., & Cohen F.E. (1997) Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci 6,1661-1681.
12. Kuhlman B., & Baker D. (2000) Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci USA 97, 10383-10388.

ROBETTA-GINZU - 211 models for 99 DP targets

Ginzu homolog identification and domain parsing in CASP7

D. Chivian[1], D. E. Kim[2], L. Malmström[2] and D. Baker[2]

[1] – Lawrence Berkeley National Laboratory, Berkeley, CA

[2] – University of Washington, Seattle, WA

DCChivian@lbl.gov

Protein chains often contain more than one domain. In order to predict the domain organization of a protein, we have developed the Ginzu[1,2] homolog identification and domain parsing method. The method is available to the public as part of the Robetta server[1,3,4] (http://robetta.org).

Ginzu attempts to determine the locations of putative domains in the query sequence and the identification of any likely homologs with experimentally characterized structures. These steps are not decoupled, since the ability to assign a region of the target to a known protein structure greatly increases the likelihood that it is at least one protein domain. The approach consists of scanning the target sequence with successively less confident methods to assign regions that are likely to be domains. Once those regions are identified, cut points in the putative linkers are determined, and if possible a single homologous PDB chain is associated with each putative domain. The initial scan attempts to identify the closest relatives with experimental structures to regions of the query sequence. A straightforward BLAST/PSI-BLAST[5] search against the PDB sequence database detects such relatives. All PDB ids that are detected at this stage are stored. Non-overlapping regions that possess the best combination of detection confidence and length of coverage are assigned as domains. The associated PDB id and region of the chain matched is retained.

One may then employ more remote fold-recognition methods to detect homologous PDB structures. We used 3D-Jury-A1[6] in this step for the parsing of the CASP7 targets. Again, as with the PSI-BLAST detections, the associated PDB and region of the target chain covered is retained.

Any remaining long regions of the query that do not have structural homologs may require further division into domains. One may search unassigned regions against Pfam[7]. Subsequent steps of Ginzu utilize the program "msa2domains", which examines the PSI-BLAST multiple sequence alignment (MSA) to find clusters of sequences in the PSI-BLAST multiple sequence alignment (MSA) and assigns these as regions of increased likelihood of possessing a domain. This is done in an order based on the number of unique observations in the cluster (essentially a non-redundant depth), with overlaps not permitted. Lastly, msa2domains determines where to place the exact cut points in the linker regions, or any remaining long unassigned regions, via a heuristic that again considers clusters of sequences in the PSI-BLAST MSA, the least

occupied positions in the MSA, strongly predicted loop regions by PSIPRED[8], and distance from the nearest region of increased domain confidence. A fourth term boosts the likelihood of a domain boundary in regions of the MSA where the sequences frequently begin or end.

The final step consists of parsing regions that have been assigned structural homologs based on the model generated by that assignment. We have developed a consensus variant of Taylor's structure-based domain parsing method[9] that is applied to the target's final Robetta model, as well as PSI-BLAST detectable structural homologs, to complete the domain parsing. Alternate domain predictions based on the model from the default K*Sync alignment to the parent are also returned, as are MSA-based predictions for weak confidence 3D-Jury detected regions.

1. Chivian D., Kim D.E., Malmstrom L., Bradley P., Robertson T., Murphy P., Strauss C.E., Bonneau R., Rohl C.A., & Baker D. (2003) Automated prediction of CASP-5 structures using the Robetta server. Proteins 53, 524-533.
2. Kim D.E., Chivian D., Malmstrom L., & Baker D. (2005) Automated prediction of domain boundaries in CASP6 targets using Ginzu and RosettaDOM. Proteins 61, 193-200.
3. Kim D.E., Chivian D., & Baker D. (2004) Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res 32, W526-W531.
4. Chivian D., Kim D.E., Malmstrom L., Schonbrun J., Rohl C.A., & Baker D. (2005) Prediction of CASP6 structures using automated Robetta protocols. Proteins 61, 157-166.
5. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., & Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25, 3389-3402.
7. Ginalski K., Elofsson A., Fischer D. & Rychlewski L. (2003) 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 19, 1015-1018.
8. Bateman A., Birney E., Cerruti L., Durbin R., Etwiller L., Eddy S.R., Griffiths-Jones S., Howe K.L., Marshall M., & Sonnhammer E.L. (2002) The Pfam protein families database. Nucleic Acids Res 30, 276-280
9. Jones D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292, 195-202.
10. Taylor W.R. (1999) Protein structural domain identification. Protein Eng 12, 203-216.

ROKKO - 476 models for 98 3D targets

Template-free Prediction by Fragment Assembly with SimFold Energy Function at CASP7

S.J. Park[1], N. Hori[2], K. Okazaki[2] and S. Takada[12]

[1] - Faculty of Sci, Kobe Univ, [2] - Grad School, Sci & Tech Kobe Univ

stakada@kobe-u.ac.jp

Team ROKKO primarily focused on predicting structures that need template-free modeling and could have previously unseen folds. Prediction method, statistics, and short description for each target are available at http://www.proteinsilico.org/ROKKO/casp7/rokko_casp7_strategy.html.

(1) General Workflow: All targets automatically stream to the general sequence analysis procedure. BLAST1 package first searches homologous templates through NRDB, and then mainly PSI-PRED2 predicts secondary structure elements (SSEs) using filtered NRDB. Some of DBs used are weekly updated. If significant templates for a target are found, all available information on the templates is gathered for selecting high-resolution structural templates (See (2)). When templates do not exist, 3D-Jury3 templates and alignments are gathered. When we did not get reliable templates, we performed fragment assembly simulations either by MCFA and/or GAFA (See (5) and (6)).

(2) Template-based Prediction: If we are satisfied with the quality of a template BLAST found, we sample template-target sequence alignments using the stochastic backtracking procedure4 (over 100 sub-optimal alignments). When several templates are covering distinct target regions, we randomly pick alignments from each ensemble of the sub-optimal alignments, and input them as initial alignments of the progressive multiple sequence alignment (approximately 1000-3000 alignments). We also use template-target profile alignments when PSI-BLAST found templates with relatively higher E-value (>0.001). We convert the alignments to 3D structures by running MODELLER5, and evaluate them using both of Verify3D6 and Prosa7 to check the initial alignment quality. We iteratively run MODELLER with seemly good alignments, and repeatedly checked SSEs and the quality of local/global structures. After ending this iterative procedure, we select final models from the 2D score distribution generated by Verify3D and Prosa.

(3) Fragment Library Construction: For template-free prediction, we first build a set of 10-residue segments by comparing the feature vector of a target sequence with them of library containing 2598 known-structure proteins that share <25% sequence identity. The vector contains PSSM of PSI-BLAST, grouped chemical property of a residue, and a SSE. Two types of fragment libraries are generated. Type I; a correlation coefficient scores top 200 segments for each overlapping 10-residue fragment of a target. Type II; five

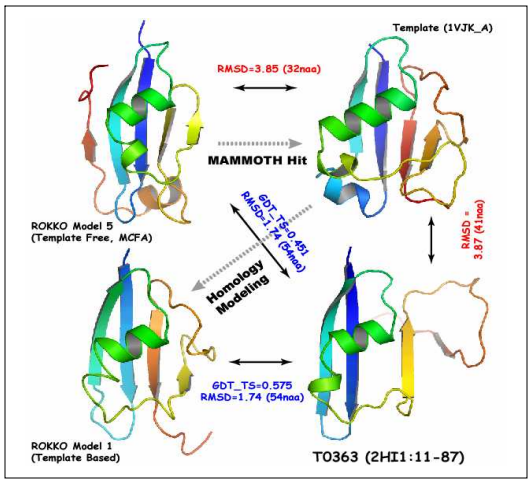
scoring functions including the correlation coefficient pick over 200 segments by considering the degree of dominated level (often called “Pareto Frontier” in multi-objective optimization field).

(4) SimFold Energy Function: For fragment assembly simulations, we solely used a coarse-grained model, SimFold8,9, in which side chain atoms are replaced with a center of mass. SimFold contains van der Waals interaction, secondary structure propensity, hydrogen bond interaction, hydrophobic interaction, and pairwise interaction. The latter three terms depend on the degree of burial of interacting atoms. No protein specific potential such as secondary structure prediction based potential is used in the energy function. Parameters in SimFold are optimized by Z-score optimization method.

(5) Multi-Canonical Ensemble Fragment Assembly (MCFA): Using Type I fragment library, we performed the reversible MCFA10 that fulfills detailed balance condition. The predictive accuracy of our MCFA in de novo prediction has been proved in CASP6. On the other hand, to define a reasonable weight function of MCFA is very time-consuming and human-dependant. We applied, therefore, Wang-Landau algorithm11 to the MCFA with a slight modification. We arranged the reducing schedule of the Wang-Landau factor by using our empirical data, and defined a weight function through approximately 2-3 billion Monte Carlo steps in a MCFA. Independent MCFA for a target sampled conformations as many as possible by the given time limit.

(6) Genetic Algorithm Fragment Assembly (GAFA): Using Type II fragment library, we test a Genetic Algorithm newly developed (its basic code came from the earlier study12). With 100 initial random coils, the GA randomly selects a residue as a crossover point. Based on this point, GA shuffles the two parents randomly selected from the current population, and replaces a segment (4-10 residues long) to a fragment coming from the library. After generating 200 offspring, GA updates the parents with the lowest-energy child and random one. Thus, the initial coils hopefully evolve to the lowest-energy conformations through 5000 update steps. The final conformations of independent GAFA runs are gathered as many as possible, and are analyzed.

(7) How We did in CASP7: We performed the template-based procedure for predicting targets that have significant PSI-BLAST E-value (< 0.001) or 3D-jury jscore (>50.0). For all remaining targets, we conducted MCFA and/or GAFA with the different types of fragment libraries. When there exist long alignment gaps (>20 residues) or probably unseen domains (e.g. T0311, T0347, etc.), we first predicted a full-length model with a template and ran FA to predict these broken regions. For a target that is likely to have multiple domains, we parsed it into monomers based on domain DBs, and combined them into a single chain by FA. 13 targets were predicted by the consensus of MCFA and GAFA; by using cluster analysis and visual inspection, we selected five models from independently sampled models by each FA method.



Interestingly, we often found that some of models FA predicted have high structural similarity to known proteins. In such cases, we added a template-based model to the final models if we were confident. For example, in T0363 case, we first selected five models from MCFA samples. MAMMOTH13 said that all five models are considerably similar to a Beta Grasp Fold. Particularly, model_5 was highly similar to 1MG4_A (z_score=4.92) that is akin to 3D-Jury templates. We believed, consequently, that 1VJK_A (jscore=46.88) is the best template for T0363. Such conspicuous structural similarity with remote homology was found from FA models of several targets (e.g. T0304, T0349, T0353, T0361, T0382, etc.). Surprisingly, a model of SimFold FA for T0383 culled 1QYN_A (jscore=6.25) that is 3.66 Angstroms over 70 residues of the T0383 native.

It is deemed again that SimFold FA is feasible to capture the native-like interactions from high quality fragment library. Therefore, the reliability of structural templates fold recognition servers detected can be confirmed to increase the predictive accuracy. This will be a steppingstone to better prediction of new folds.

1. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W. & Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389-3402.
2. Jones D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292, 195-202.
3. Ginalski K., Elofsson A., Fischer D. & Rychlewski L. (2003) 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 22, 1015-1018.
4. Muckstein U., Hofacker I.L., & Stadler P.F. (2002) Stochastic pairwise alignments. Bioinformatics 18, S153–S160.
5. Sali A. & Blundell T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234, 779-815.
6. Bowie J.U., Luthy R. & Eisenberg D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. Science 253, 164-170.
7. Sippl M.J. (1993) Recognition of errors in three-dimensional structures of proteins. Proteins 17, 355-362.

8. Fujitsuka Y., Takada S., Luthey-Schulten Z.A. & Wolynes P.G. (2004) Optimizing physical energy functions for protein folding. Proteins 54, 88-103.
9. Fujitsuka Y, Chikenji G, & Takada S. (2006) SimFold energy function for de novo protein structure prediction: Consensus with Rosetta. Proteins 62, 381-398.
10. Chikenji G., Fujitsuka Y. & Takada S. (2003) A reversible fragment assembly method for de novo protein structure prediction. J. Chem. Phys. 119, 6895-6903.
11. Wang F. & Landau D.P. (2001) Efficient, multiple-range random walk algorithm to calculate the density of states. Phys. Rev. Lett. 86, 2050–2053.
12. Park S.J. (2005) A study of fragment-based protein structure prediction: biased fragment replacement for searching low-energy conformation. Genome Informatics 16, 104-113.
13. Ortiz A.R., Strauss ,C.E. & Olmea O. (2002) MAMMOTH (Matching molecular models obtained from theory): An automated method for model comparison. Protein Science 11, 2606-2621.

ROKKY - 444 models for 98 3D targets

De novo Structure Prediction Server by Fragment Assembly with SimFold Energy Function

W. Jin[1], S.J. Park[1], and S. Takada[12]

[1] – Faculty of Sci, Kobe Univ, [2]- Grad School, Sci & Tech Kobe Univ

stakada@kobe-u.ac.jp

ROKKY is a fully automated server that predicts protein structures from a given amino acid sequences with/without templates. It primary emphasizes the template-free targets by using Simulated Annealing Fragment Assembly (SAFA) with SimFold[1-4], a coarse-grained physico-chemical energy function. Although the predictive accuracy of ROKKY in template-free predictions was highly evaluated at CASP6, we slightly modified its job flow for CASP7 targets.

Here, we briefly describe (1) job flow, (2) generation of fragment candidates, and (3) SAFA with SimFold and model selection.

1) Job flow: For all targets, ROKKY first performs PSI-BLAST[5] using NR and PDB, respectively. When templates with e-value smaller than 0.001 is found, ROKKY uses its alignment and makes model structures by running MODELLER[6]. Otherwise, ROKKY submits the target sequence to 3D-Jury meta-server[7] and obtains the results. When templates (3D-Jury score > 50.0) are found, ROKKY uses 3D-Jury's templates and alignments. MODELLER also constructs variable loops if exist as alignment gaps. For the rest, ROKKY performs SAFA with SimFold energy function for parts of the unaligned sequence that is longer than 30 residues and choose 5 models in the sampled structures based on clustering analysis. For multi-domain targets, individually modeled domains are docked to have a model of the whole sequence by SAFA.

2) Generation of fragment candidates: For every 10-residue in the query sequence, the correlation coefficient of 20×10 dimensional fragment vectors made of the PSSM from PSI-BLAST retrieves fragment candidates from 2600 template proteins that have known structures. The collection of 50 fragment candidates for each site of the target overlapping is filtered by Ramachandran plot if PSI-PRED[8] predicted the site is helix with high confidence.

3) Fragment assembly (FA) with SimFold and model selection: ROKKY performs SAFA with SimFold using fragment candidates generated by (2) for the targets or domains that has no apparent templates. SA algorithm replaces a randomly chosen fragment (4-9 residues) with another fragment randomly chosen from the candidates by following Metropolis judgment. Selection temperature is gradually decreased to obtain low-energy structures. This SAFA runs are repeated as many samples as possible till a few hours to the time deadline. The samples that have secondary structure more than a certain cutoff are treated by the cluster analysis with the group average method, in which the centers of the five largest clusters are chosen as final models.

Computational resources of ROKKY are partially provided by Human Genome Center of University of Tokyo.

1. Takada S (2001) Protein folding simulation with solvent-induced force field: folding pathway ensemble of three-helix-bundle Proteins. Proteins 42, 85-98.
2. Fujitsuka Y., Takada S., Luthey-Schulten Z.A. & Wolynes P.G. (2004) Optimizing physical energy functions for protein folding. Proteins 54, 88-103
3. FujitsukaY., Chikenji G. & Takada S. (2006) SimFold energy function for de novo protein structure prediction: Consensus with Rosetta. Proteins 62, 381-398.
4. Chikenji G., Fujitsuka Y. & Takada S. (2003) A reversible fragment assembly method for de novo protein structure prediction. J.Chem.Phys. 119, 6895-6903.
5. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W. & Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389-3402.
6. Fiser A., Do R.K. & Sali A. (2000) Modeling of loops in protein structures. Protein Science 9, 1753-1773.

7. Ginalski K., Elofsson A., Fischer D., Rychlewski L. (2003) 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 22, 1015-1018.
8. Jones D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292, 195-202.

SAM-T06 - 682 models for 100 3D/93 RR/ 7 TR targets

SAM-T06: Full 3D predictions from UCSC

Kevin Karplus[1], George Shackelford[1], Firas Khatib[1], Martin Madera[1], Grant Thiltgen[1], Zack Sanborn[1], Chris Wong[1], Pinal Kanabar[1], Cynthia Hsu[1], Crissan Harris[1], Sylvia Do[1], NavyaSwetha Davuluri[1]
[1]*-UCSC*
karplus@soe.ucsc.edu

The SAM-T06 hand predictions use methods similar to SAM-T04 in CASP6 and the SAM-T02 method in CASP5.

We start with a fully automated method, implemented as the SAM_T06 server. The server runs the SAM-T2K and SAM-T04 iterative methods for finding homologs of the target and aligning them. (The hand method also uses the experimental new SAM-T06 alignment method, which we hope is both more sensitive and less prone to contamination by unrelated sequences.) We use the resulting alignments to make local structure predictions with our neural nets. Currently we use 10 local-structure alphabets: DSSP, STRIDE, STR2 (an extended version of DSSP that splits the beta strands into multiple classes: parallel / antiparallel / mixed, edge / center), ALPHA (a discretization of the alpha torsion angle between CA(i-1), CA(i), CA(i+1) and CA(i+2)), BYS (a discretization of Ramachandran plots due to Bystroff), CB_burial_14_7 (a 7-state discretization of the number of C_beta atoms in a 14A radius sphere around the C_beta), near-backbone-11 (an 11-state discretization of the number of residues in a 9.65A radius sphere around a residue), DSSP_EHL2 (CASP's collapse of the DSSP alphabet; computed as a weighted average of the other backbone alphabet predictions), O_NOTOR2 (an alphabet for predicting characteristics of hydrogen bonds from the carbonyl oxygen) and N_NOTOR2 (an alphabet for predicting characteristics of hydrogen bonds from the amide nitrogen).

The server makes two-track HMMs with each alphabet (a weight of 1.0 for the amino-acid track and 0.3 for local structure) and uses them to score a library of about 8000 (t06), 10000 (t04), or 15000 (t2k) templates. The template libraries are expanded weekly, but old template HMMs are not rebuilt. We also used a single-track HMM to score not just the template library, but a non-redundant copy of the entire PDB.

One-track HMMs built from the template library multiple alignments were used to score the target sequence. All the logs of e-values were combined in a weighted average (with rather arbitrary weights, since we still have not taken the time to optimize them), and the best templates ranked.

Alignments of the target to the top templates were made using several different alignment methods (mainly using the SAM hmmscore program, but a few alignments were made with Bob Edgar's MUSCLE profile-profile aligner). Generate fragments (short 9-residue alignments for each position) using SAM's "fragfinder" program and the 3-track HMM which tested best for alignment.

Residue-residue contact predictions are made using mutual information, pairwise contact potentials, joint entropy, and other signals combined by a neural net.

Then the "undertaker" program (named because it optimizes burial) is used to try to combine the alignments and the fragments into a consistent 3D model. No single alignment or parent template was used as a frozen core, though in many cases one had much more influence than the others. The alignment scores were not passed to undertaker, but were used only to pick the set of alignments and fragments that undertaker would see. Helix and strand constraints generated from the secondary-structure predictions are passed to undertaker to use in the cost function, as are the residue-residue contact prediction.

One important change in this server over previous methods is that sheet constraints are extracted from the top few alignments and passed to undertaker.

After the automatic prediction is done, we examine it by hand and try to fix any flaws that we see. This generally involves rerunning undertaker with new cost functions, increasing the weights for features we want to see and decreasing the weights where we think the optimization has gone overboard. Sometimes we will add new templates or remove ones that we think are misleading the optimization process.

New this year, we are also occasionally using ProteinShop to manipulate proteins by hand, to produce starting points for undertaker optimization. We expect this to be most useful in new-fold all-alpha proteins, where undertaker often gets trapped in poor local minima by extending helices too far.

Another new trick is to optimize models with gromacs to knock them out of a local minimum. The gromacs optimization does terrible things to the model (messing up sidechains and peptide planes), but is good at removing clashes. The resulting models are only a small distance from the pre-optimization

models, but score much worse with the undertaker cost functions, so undertaker can move them more freely than models it has optimized itself.

SAM-T06 - 682 models for 100 3D/93 RR/ 7 TR targets

Residue-Residue Contact Prediction Using Selected Correlation Statistics

George Shackelford[1] and Kevin Karplus[1]
[1] - *University of California, Santa Cruz*
ggshack@soe.ucsc.edu

We present a neural network based residue-residue predictor using selected statistics. When we were developing the CASP6 predictor, we used every input we thought might be useful even when they may be redundant. Neural networks can still learn when there are redundant inputs but the learning is usually not as effective with respect to predictions. For our new network, we conduct a series of experiments to determine a more effective set of inputs.

The primary source of data is a multiple sequence alignment provided by SAM-T04. We assume correlated mutations as a significant indication of contact, therefore we consider a variety of correlation statistics over the i and j columns as possible inputs. The two we finally found the most effective are one: an e-value based on mutual information and two: a propensity statistic using the sum of log propensities of residue pairs between the two columns.

Besides those two inputs we use local structure predictions and residue distributions also provided by SAM-T04. The inputs for the correlation statistics are based on columns from thinning the MSA to 50 precent, i.e., sequences are removed from the MSA until no two sequences have more than the 50 percent sequence identity. We use residue distributions for columns adjusted by Derlich mixtures. Windows around i and j are frequently used, e.g., a window of five around i and j are the columns in i-2 to i+2 and in j-2 to j+2.

The 449 inputs we finally use are:

- The two corrrelation statistics mentioned above.

- log(sequence length) , log(separation),

- window of five of the distributions and their respective entropy,

- window of five for two local structure predictions: one predicts secondary structure using a superset of DSSP and the other predicts how deeply buried the residue is from the surface of the protein.

Other correlation statistics tested but not used in the predictor included raw mutual information, mutual information over entropy, BASC, OMES (Observed Minus Expect, Squared). In training the neural network we find that Improved Resilient Backpropagation helps in convergence.

The resulting predictor shows improvements over the CASP6 predictor.

PROTINFO - 500 models for 100 3D targets
SAMUDRALA - 611 models for 99 3D /5 FN /99 QA/ 4 TR targets

Comparative model refinement using graph-theoretic and consensus-based restrained molecular dynamics approaches (PROTINFO/SAMUDRALA)

T. Liu, L-H Hung, S-C. Ngan and R. Samudrala
University of Washington
ram@compbio.washington.edu

We developed and evaluated two novel methods for refining template-based predictions at CASP7. Initial models were generated based on alignments provided by the 3D-Jury server (http://bioinfo.pl/meta) [1] using our protein structure modeling server, PROTINFO (http://protinfo.compbio.washington.edu) [2, 3]. Additional initial models were obtained from the CAFASP5 server after scrutinizing the alignments to gain extra variability in sequence alignments and templates. We then refined initial models using two methods: A graph-theoretic clique finding (CF) approach and a restrained molecular dynamics simulations using consensus-based constraints. The latter is recently developed to address the refinement problem in CASPR and in CASP7. The models generated using both approaches were minimised using ENCAD [4] to produce good geometry and packing, and the five best scoring models were submitted as our CASP7 predictions.

The CF approach has been developed to handle the large conformational space of main chain and side chain possibilities resulting from the interconnected nature of interactions in protein structures [5]. Our enhanced refinement method employed the CF approach in a fully automated manner to mix and match regions between different initial models for a given target protein. Sampling of side chain and main chain conformations was accomplished by exhaustively enumerating all possible choices from a population of initial models. The best combinations of these possibilities were selected through a graph-theoretic clique finding approach aided by our all-atom conditional probability discriminatory function (RAPDF) [6]. This process typically

generates an optimized conformation ensemble representing the best combination of secondary structures, resulting in the refined models of higher quality.

For the second refinement method, consensus distance constraints and dihedral angles were compiled from the initial models and structures were generated by restrained molecular dynamics simulations using the software CYANA (Combined Assignment and Dynamics Algorithm for NMR Applications, © by Peter Güntert) [7]. Distances between two non-local atoms (separated by more than four residues) were measured and binned in 0.5 Å increments. Atom pairs within a distance bin observed in all the best scoring initial models were considered consensus distance restraints. The values of upper and lower limits for each such restraint were determined by the observed distances in the models. Each consensus distance restraint was ranked based on its RAPDF score for individual atom pairs [6]. The highest ranked consensus restraints were considered to be more accurate and were used for the restrained molecular dynamics accuracy. For the calculation of dihedral angle restraints, Ψ/Φ angles of consensus residues from the 3D-Jury alignment were used. The distance and angle restraints were then directly input to CYANA which generated a set of conformations satisfying the input restraints using torsion angle dynamics. Although this method is still in its early stage of development, it represents our first attempt to move a template-based model closer to its native fold. Preliminary analyses of targets for which the experimental structures have been released, and of targets in the CASPR experiment, indicate that in some cases our protocol is able to produce significant (> 1Å) refinements relative to the starting model.

1. Ginalski K., Elofsson A., Fischer D. & Rychlewski L. (2003) 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics. 19: 1015-1015.
2. Hung L.H. & Samudrala R. (2003) PROTINFO: Secondary and tertiary protein structure prediction. Nucleic Acids Research 31: 3296-3299.
3. Hung L.H., Ngan S.C., Liu T. & Samudrala R. (2005) PROTINFO: New algorithms for enhanced protein structure prediction. Nucleic Acids Research 33: W77-W80.
4. Levitt M., Hirshberg M., Sharon R. & Daggett V. (1995) Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. Comp Phys Comm 91: 215-231.
5. Samudrala R. & Moult J. (1998) A graph-theoretic algorithm for comparative modeling of protein structure. J Mol Biol 279: 287-302.
6. Samudrala R. & Moult J. (1998) An all-atom distance dependent conditional probability discriminatory function for protein structure prediction. J Mol Biol 275: 893-914.
7. Güntert P., Mumenthaler C. & Wüthrich K. (1997) Torsion angle dynamics for NMR structure calculation with the new program DYANA. J. Mol. Biol. 273: 283-298.

Samudrala-AB - 493 models for 99 3D targets

Constraint-based free modeling

L-H. Hung[1], T. Liu[1], S-C. Ngan[1] and R. Samudrala[1]

[1] – *University of Washington*

lhhung@compbio.washington.edu

PROTINFO-AB/SAMUDRALA-AB are free modeling methodologies that do not use template coordinates. The two methods differ mainly in the amount of computation time used, the starting models used for refinement, and some minor changes in implementation that occurred as CASP7 progressed. Initial starting models are obtained from 3D-Jury (http://bioinfo.pl/meta)[1] when there is a significant match and/or from CASP server models. These models are then ranked using SAMUDRALA-MQAP. The highest ranking models are used to derive constraints using a consensus approach and CYANA[2] is used to generate models satisfying those constraints, in a procedure similar to that used for PROTINFO/SAMUDRALA. Models are then ranked on the basis of our all atom energy function (RAPDF)[3], hydrogen bonding and iterative clustering[4].

The variable regions of the best models are then rebuilt *de novo* using a Monte Carlo simulated annealing search procedure. The move sets used are continuous phi/psi distributions derived from experimentally determined structures. The target that is minimised is a combination of our all-atom energy function (RAPDF), a hydrophobic compactness function (HCF), and a function that penalises bad contacts[4-7]. In the absence of good starting models, the entire protein is simulated *de novo* using this procedure. The best scoring models are then used to obtain a second set of constraints which are used by CYANA to generate a new set of models. The final set of models is then scored again and the five best are submitted.

Preliminary analysis of CASP7 targets for which the experimental structure has been released indicates that our approach complements PROTINFO/SAMUDRALA (i.e., there are several targets for which one approach produces excellent models and the other does not, and *vice versa*). In some cases, particularly for harder targets, the models produced are of higher accuracy than any available template.

1. Ginalski K., Elofsson A., Fischer D. & Rychlewski L. (2003) 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics. 8, 1015-1018.
2. Guntert P. (2004) Automated NMR structure calculation with CYANA Methods Mol Biol. 278, 353-378.
3. Samudrala R. & Moult J. (1997) An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J. Mol. Biol. 275, 893-914,.
4. Hung L-H., Ngan S-C., & SamudralaR. (2007) Ab initio protein structure prediction. Computational Methods for Protein Structure Prediction and Modeling 2 (in press).
5. Hung L-H. & Samudrala R. (2006) Accurate structures from sparse and noisy NOE constraints using continuous dihedral angle distributions and RAPDF-based target functions. submitted.
6. Hung L-H., Ngan S-C., Liu T. & Samudrala R. (2005) PROTINFO: New algorithms for enhanced protein structure prediction. Nucleic Acids Res. 33, W77-W80..
7. Hung L-H. & Samudrala R. (2003) PROTINFO: Secondary and tertiary protein structure prediction. Nucleic Acids Res. 31, 3296-3299.

SBC - 500 models for 100 3D targets

Automatic predictions of protein structure, quality assessments, local residue based quality and function from Stockholm University.

B. Wallner[1,2], D. Ekman[1,2], Å.K. Björklund[1,2] and A. Elofsson[1,2]

1Stockholm Bioinformatics Center, 2 Center for Biomembrane Research, Stockholm University 106 91 Stockholm, Sweden
arne@bioinfo.se

For CASP7 we have applied a number of automated methods to predict the structure, quality and function of the CASP targets. The predictions are based on published and freely available methods developed during the last years in our group. Below follows a short description of our automatic prediction methods as well as references to the papers and websites containing more detailed information. All prediction information is available at: http://www.pdc.kth.se/~bjornw/casp7/targets/

Structure Prediction

We have submitted three structure prediction methods, Pcons6, Pmodeller6 and SBC. The two first ones were submitted as server prediction while SBC was submitted as a manual prediction, but no manual interference was used. The Pcons6 method is a "consensus" methods identical to Pcons51, where the similarity of models collected by an in-house developed meta-serve, http://www.cbr.su.se/pcons/, is compared. The meta-server tried to use the following methods: samt02, blast, robetta, bas_b, bas_c, ffas03, orfeus, pdbblast, mgenthreader, blast, mbam, forte, sp3, orfbc, fugsa and inbgu. Pmodeller6 is simple an approach to from all methods that have a Pcons6 score within 30% from the best score choose the model that have the highest ProQ score2. The SBC methods is identical to the Pmodeller6 method but uses all server predictionsubmitted to CASP as an input.

Quality assessment methods.

Four quality assessment methods were applied in CASP7, ProQ , Pcons, ProQprof and ProQlocal. All these method predicts the quality for each residue as well as for the entire model3. The ProQ QA method is based on a neural network trained on structural features, ProQprof uses sequence similarity and ProQlocal is a combination of these two methods. The Pcons QA method is based on the local structural similarity between all models submitted to CASP7. ProQ is available as a webserver at http://www.sbc.su.se/~bjornw/ProQ/ProQ.cgi and the local quality predictor is available at http://sbcweb.pdc.kth.se/cgi-bin/bjornw/ProQres.cgi

Function prediction methods.

We have developed two simple function prediction methods, SBCdomfun, SBCseqfun. SBCdomfun, http://sbcweb.pdc.kth.se/cgi-bin/diaek/domsearch.cgi, is based on mutual information between GO terms and Pfam domains. The Pfam domains were detected using profile-profile searches at the http://bioinfo.pl meta-server. SBCseqfun, http://sbcweb.pdc.kth.se/cgi-bin/diaek/seqfunction.cgi , is based on searches against a sequence database with annotated sequence. The most frequent GO-terms among the top-hits were used.

Automated analysis of structure prediction methods

As of Sept 28 we have performed an automated analysis of all server predictions as well as some manual predictions for the 81 CASP targets solved at this date. These results show that Pmodeller6 is the second best server overall although it does not perform very well on the "easy" targets while SBC is second only to the clearly method, Zhang-server. Here follows a list of selected results for the best

methods available at http://www.pdc.kth.se/~bjornw/casp7/targets/results/

Score and (rank of automatic methods):

Method	All	EASY	HARD

Zhang-Server	50.52 (1)	34.84 (1)	15.68 (1)
SBC	48.66	33.50	15.15
Pmodeller6	47.31 (2)	32.59 (24)	14.72 (2)
HHpred2	47.11 (3)	32.94 (15)	14.17 (4)
Robetta	47.06 (4)	32.62 (23)	14.44 (3)
CIRCLE	47.03 (5)	33.41 (4)	13.62 (7)
Pcons6	46.86 (6)	32.84 (18)	14.03 (6)
UNI-EID_expm	46.61 (7)	33.50 (2)	13.10 (14)
beautshot	46.56 (8)	33.40 (5)	13.16 (13)
FAMSD	46.55 (9)	33.26 (6)	13.29 (12)
MetaTasser	46.41 (11)	32.31 (27)	14.10 (5)
shub	45.71 (18)	33.41 (3)	12.30 (23)

1. Wallner B. and Elofsson A. (2003) Can correct protein models be identified Protein Science 12(5):1073-86
2. Wallner B. and Elofsson A. (2005) Pcons5: combining consensus, structural evaluation and fold recognition scores. Bioinformatics 21(23):4248-54
3. Wallner B. and Elofsson A. (2006) Can correct regions in protein models be identified. Protein Science 15(4):900-13.

SCFBio-IITD - 20 models for 3 3D targets

Bhageerath: An Energy Based Protein Tertiary Structure Prediction Server for Small Globular Proteins.

B. Jayaram, Kumkum Bhushan, Lipi Thukral,
Sandhya R. Shenoy, Pooja Narang, Surojit Bose,
Praveen Agrawal, Debashish Sahu and Vidhu Pandey

Department of Chemistry &
Supercomputing Facility for Bioinformatics and Computational Biology,
Indian Institute of Technology Delhi,
Hauz Khas, New Delhi 110016, India
bjayaram@chemistry.iitd.res.in

The tertiary structure prediction of a protein using the amino acid sequence information alone is one of the fundamental unsolved problems in computational biology[1]. Significant progress has been made in recent years in generating computational solutions based on laws of physics. This approach, commonly referred to as ab initio[2-4] is based on the thermodynamic hypothesis formulated by Anfinsen, according to which the native structure of a protein corresponds to the global minimum of its free energy under given conditions[5]. Protein structure prediction using de novo method is accomplished by a search for a conformation corresponding to the global-minimum of an appropriate potential energy function without the use of secondary structure prediction, homology modeling, threading etc.[6].

We describe here an energy based computer software suite for narrowing down the search space of tertiary structures of small globular proteins. The protocol comprises eight different computational modules that form an automated pipeline. It combines physics based potentials with biophysical filters to arrive at 10 plausible candidate structures starting from sequence andsecondary structure information. The methodology has been validated here on 50 small globular proteins consisting of 2-3 helices and strands with known tertiary structures. For each of these proteins, a structure within 3-6 Å RMSD(root mean square deviation) of the native has been obtained in the 10 lowest energy structures. The protocol has been web enabled and is accessible at http://www.scfbio-iitd.res.in/bhageerath.

The first module involves the formation of a three-dimensional structure from the amino acid sequence with the secondary structural elements as input from the user. For CASP7 targets, the secondary structure information was taken from online server GORV[7-8] and an in-house developed program named PROSECSC. The second module involves generation of a large number (~10^5 to 10^6) of trial structures with a systematic sampling of the conformational space of loop dihedrals. The trail structures thus generated are screened in the

third module through the persistence length and radius of gyration filters[9], developed for the purpose of reducing the number of improbable candidates. The resultant structures are refined in the fourth module by a Monte Carlo method to remove steric clashes and overlaps involving atoms of main chain and side chains. In module five, the structures are energy minimized to further optimize the side chains. Module six involves ranking of structures using an all atom energy based empirical scoring function[10] followed by a selection of the 100 lowest energy structures. Module seven reduces the probable candidates based on an index developed using the regularity observed in protein loop dihedrals. Module eight further reduces the structures selected in the previous module to 10 using topological equivalence criterion and accessible surface areas. The above eight modules are configured to work in conduit in an automated mode. Further refinement of structures was carried out by Molecular dynamics studies and subsequent energy scans for CASP7 to reduce the final number to five.

The webserver (Bhageerath; www.scfbio-iitd.res.in/bhageerath) and the automated computational protocol developed and embedded in the software for a prediction of the three dimensional structures of small proteins is described.

1. Liwo A., Khalili M. and Scheraga H.A. (2005) Ab initio simulation of protein-folding pathways by molecular dynamics with united residue model of polypeptide chains. Proc. Natl. Acad. Sci. USA, 102, 2362-2367.
2. Scheraga H.A. (1992) Some approaches to the multiple-minima problem in the calculation of polypeptide and protein structures. Int. J. Quant. Chem., 42, 1529-1536.
3. Scheraga H.A. (1996) Recent developments in the theory of protein folding: searching for the global energy minimum. Biophys. Chem., 59, 329-339.
4. Vasquez M., Nemethy G. and Scheraga H.A. (1994) Conformational energy calculations on polypeptides and proteins. Chem. Rev., 94, 2183.
5. Anfinsen C.B. (1973) Principles that govern the folding of protein chains. Science, 181, 223.
6. Pillardy J. (2001) Recent improvements in prediction of protein structure by global optimization of a potential energy function. Proc. Natl. Acad. Sci. USA, 98, 2329-2333.
7. Kloczkowski A., Ting K.-L, Jernigan R.L., Garnier J. (2002) Combining the GOR V algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence. Proteins, 49, 154-166.
8. Sen T.Z., Jernigan R.L., Garnier J., Kloczkowski A. (2005) GOR V server for protein secondary structure prediction, Bioinformatics, 21(11), 2787-2788.
9. Narang P., Bhushan K., Bose S. and Jayaram B. (2005) A computational pathway for bracketing native-like structures for small alpha helical globular proteins. Phys. Chem. Chem. Phys., 7, 2364-2375.
10. Narang P., Bhushan K., Bose S. and Jayaram B. (2006) Protein structure evaluation using an all-atom energy based empirical scoring function. J. Biomol. Str. Dyn. 23 (4), 385-406.

Scheraga - 220 models for 43 3D targets

Physics-based protein-structure prediction using mesoscopic dynamics and the Conformational Space Annealing (CSA) method with the UNRES force field - test on CASP7 targets

C. Czaplewski[1,2], S. Ołdziej[1,2], M. Chinchio[1], A.V. Rojas,[1,3,4] R. Kaź mierkiewicz,[1,2] Y.A. Arnautova[1], J.A. Vila[1,5], M. Khalili[1], R.K. Murarka,[1] A. Spasic,[1] H. Shen,[1] A. Liwo[1], and H.A. Scheraga[1*]

[1] – Baker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca, NY, 14853-1301, [2] – Faculty of Chemistry, University of Gdańsk, ul. Sobieskiego 18, 80-952 Gdańsk, Poland , [3] - Department of Physics and Astronomy, Louisiana State University, Baton Rouge, LA 70803-4001, [4] – Center for Computation and Technology, Louisiana State University, Baton Rouge, LA 70803-4001, [5] – IMASL-CONICET, Facultad de Ciencias Fisico Matematicas y Naturals, Universidad Nacional de San Luis, Ejercito de los Andes 950, 5700 San Luis, Argentina

[*] has5@cornell.edu

The structures of the target proteins were predicted using our hierarchical approach[1] in which a polypeptide chain is initially treated at a united-residue level using our UNRES force field, and the coarse-grained structures thus found are subsequently converted to all-atom structures.

In the UNRES model, the atoms of the peptide group and side chain of each amino-acid residue are replaced with two centers of interactions: the united peptide group (p) located in the middle between two consecutive α-carbon atoms and the united side chain (SC). The lengths of the virtual $C^{\alpha}\ldots C^{\alpha}$ and $C^{\alpha}\ldots$SC bonds are held fixed, but the virtual-bond angles, the virtual-bond dihedral angles, and the orientations of the $C^{\alpha}\ldots$SC virtual bonds are variable. The interactions of this simplified model are described by the UNRES potential derived from the generalized cumulant expansion of a restricted free energy (RFE) function of polypeptide chains[1]. The cumulant expansion enabled us to determine the functional forms of the multibody terms in UNRES. The potential was optimized by applying our novel hierarchical optimization method targeted at decreasing the energy while increasing the native-likeness of structures of the training proteins[2].

We used our two techniques to search the conformational space: the conformational space annealing (CSA) method and molecular dynamics which was recently introduced to UNRES[3] enhanced with multiplexing replica exchange (abbreviated MREMD);[4] this MD approach is still under development and was used only for smaller α-helical proteins. The second technique enabled us to select models based on thermodynamic stability of the calculated ensembles. To speed up the search for larger proteins, information from secondary structure prediction by PSIPRED[5] was used in the generation of the initial structures; however, the search was carried out in an unrestricted manner with the UNRES energy function. For very large α-helical proteins, a search with our simplified approach[6] in which α-helices are represented as cylinders was carried out and, for the lowest-energy structures thus obtained, the conformational search was completed with the UNRES force field.

To select final models, the conformations from CSA calculations were clustered and the families ranked according to UNRES energies. The models were selected as the lowest-energy representatives of the five lowest-energy families. The MREMD ensembles were processed by histogram reweighting to calculate the probabilities of conformations and clustered at the folding temperature (located by inspection of the calculated heat-capacity curves), and the free energy of each cluster was evaluated. The five models were chosen as average conformations from the five clusters with the lowest free energies.

1. Scheraga H.A. et al. (2004) The protein folding problem: global optimization of force fields. Frontiers in Bioscience 9, 3296-3323.
2. Ołdziej S. et al. (2004) Optimization of the UNRES force field by hierarchical design of the potential-energy landscape. 3. Use of many proteins in optimization. J. Phys. Chem. B 108, 16950-16959.
3. Khalili M. et al. (2005) Molecular dynamics with the united-residue model of polypeptide chains. I. Lagrange equations of motion and tests of numerical stability in the microcanonical mode. J. Phys. Chem. B 109, 13785-13797.
4. Nanias M. et al. (2006) Replica exchange and multicanonical algorithms with the coarse-grained united-residue (UNRES) force field. J. Chem. Theory and Comput. 2, 513-528.
5. McGuffin L.J. et al. (2000) The PSIPRED protein structure prediction server. Bioinformatics 16, 404-405.
6. Nanias M. et al. (2003) Packing helices in proteins by global optimization of a potential energy function. Proc. Natl. Acad. Sci. USA. 100, 1706-1710.

Schomburg-group - 133 models for 18 3D/65 QA targets

A comparative modeling pipeline combined with a statistical potential scoring function

P. Benkert[1] and D. Schomburg[1]

1 – Cologne University BioInformatics Center

pbenkert@uni-koeln.de, D.Schomburg@uni-koeln.de

Our comparative modeling pipeline consists of a PDB-BLAST-like protocol for parent detection1, a profile-profile alignment step, manual model building including a semi-automatic loop modeling procedure and a statistical potential for final model selection.

Parent detection is performed by the following PDB-BLAST protocol: 4 PSI-BLAST2 iterations on NCBI's non-redundant sequence database (clustered at 90% sequence identity) with E-value cut-off 0.001 followed by 1 iteration on pdbaa. One or several templates are selected manually based on the observed sequence identity to the target and the quality of the template (i.e. resolution, target coverage).

Several alternative target-template alignments are generated using a modified version of the profile-profile alignment functionality included in the Align-package, a C++ library provided by the Tosatto group3. Profiles are generated by PSI-BLAST (5 iterations on nr clustered at 90%, E<0.001). Alternative alignments are generated by applying different (sub-)optimal gap open and gap extension penalties.

Based on the template structures and the alignments, raw models are generated automatically which are then subjected to the knowledge-based loop prediction procedure. Boundaries of loop regions are determined manually using a consensus of PROFphd4, PSIPRED5 and SSpro6. Loops are retrieved from a fragment database storing fragments of the length 3-20 residues based on a PISCES7 selections (95% sequence identity, resolution < 2.5A). Loops are ranked according to the energy function described below. Since loop prediction is not fully automated yet, loops are selected manually from the top ranking loops.

Loop ranking and model quality assessment is done by a statistical potential consisting of a solvation term, a torsion potential over 3 adjacent residues and pairwise Cb-Cb potential combined with a term accounting for the agreement between predicted and observed secondary structure of the target and the model respectively. Side-chains are predicted by SCWRL8.

1. Rychlewski L., Jaroszewski L., Li W. & Godzik A. (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence

information. Protein Science 9(2):232-241.
2. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W. & Lipman D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389-3402.
3. http://protein.cribi.unipd.it/align/download.shtml
4. Rost B. (2005). How to use protein 1D structure predicted by PROFphd. In Walker, J.E. (Ed.). The Proteomics Protocols Handbook, , Totowa, NJ Humana, pp. 875–901.
5. Jones D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292, 195-202.
6. Cheng J., Randall A., Sweredoski M., Baldi P. (2005) SCRATCH: a Protein Structure and Structural Feature Prediction Server, Nucleic Acids Research,Web Server Issue, vol. 33, 72-76.
7. Wang G. and Dunbrack R.L.Jr. (2003) PISCES: a protein sequence culling server. Bioinformatics, 19:1589-1591.
8. Canutescu A.A., Shelenkov A.A. and Dunbrack R.L.Jr. (2003) A graph theory algorithm for protein side-chain prediction. Protein Science 12, 2001-2014.

SHORTLE - 401 models for 91 3D/5 TR targets

Homology modeling with Atom-Based Statistical Potentials and a Simple Genetic Algorithm

Q. Fang and D. Shortle
The Johns Hopkins University School of Medicine
Baltimore, MD 21205 USA
dshortl1@jhmi.edu

Our method of homology modeling in CASP7 was based on a distance-dependent atom-pair potential developed and optimized for modeling the energetic of atomic interactions in native proteins[1]. Previous work has shown that conformational search using a genetic algorithm (GA) method and a scoring function consisting of this atom-pair potential plus an orientation-dependent backbone hydrogen bonding potential used by ROSETTA[2] and a statistical solvation potential based on the solvent exclusion model of Lazaridis and Karplus[3] is able to efficiently refold small proteins that have been unfolded by changing every phi and psi angle by either +/- 3, 5, 7 degrees[1]. In that study, a strong correlation was found between the correctness of the structure, measured by Cα Distance Matrix Error (Cα-DME) to the native state, and the radius of gyration for low energy structures.

For targets established to be homology modeling challenges, templates were identified by PSI-BLAST and 3D-Jury of BioInfoBank Meta Server (http://bioinfo.pl/meta/). The optimal sequence alignment was inferred by comparing both PSI-BLAST and 3D-Jury outputs, and in some cases, more than one alignment or more than one template was used. Overlapping segments of the target sequence containing a single turn/loop plus the two flanking helices/strands were constructed de novo by recombination of overlapping 5- to 8- residue oligomers obtained from high resolution crystal structures on the basis of low local interaction energies[4]. Approximately 2000 structures that could be reasonably superposed on the helices/strands of the template were saved for each segment. Starting with the template, each turn/loop was replaced with a randomly selected fragment using the cyclic coordinate descent (CCD) algorithm[5] at randomly selected sites within the helix/strand of the template. To avoid significant changes in backbone structure within the turn/loop, a limit of 10 steps and a maximum of 5 degrees' change in each phi/psi angle were imposed on the CCD insertion. Up to 1000 structures were generated for each target, with each turn/loop fragment being allowed to contribute to no more than three accepted structures.

In the second step, an initial population of 300 full-length structures was selected for the genetic algorithm using a scoring function consisting of the sum of z-scores of the atom-based statistical potentials, hydrogen bonding energy and solvation energy. Conformational search proceeded by selection of two structures at random and recombination across a randomly chosen peptide bond. Side-chain minimization was carried out by two passages through a grid search of side-chain rotamers, using the penultimate library of Lovell et al[6]. When a recombinant had a score lower than the mean value from the previous generation, it was saved until 300 additional recombinant structures were generated. The same scoring function was used to select which 300 structures out of the 600 would survive in the next generation. The genetic algorithm was run for 20 generations, and 6 independent runs were conducted for each target, yielding a total of 1800 models.

In the final step, the 600 structures with the lowest atom-pair potentials were first selected, followed by a selection for the smallest radius of gyration. For most targets, the structure submitted as model 1 had the smallest RG (highest atom density), whereas the other models were chosen based on other energy terms.

At the completion of CASP7, we realized that for all targets predicted, much of the aligned template structure had been inadvertently retained, with most of the structural variation being confined to loops having fixed points of attachment to the template. Subsequent work has shown that when the aligned segments of secondary structure are allowed to move significantly, the genetic algorithm runs much more slowly and, for several templates, can only occasionally re-position the secondary structure to the same accuracy (CA-DME) as that of the template. Future success with this approach will probably require greater levels of global structural similarity in the initial population.

1. Fang Q, Shortle D. (2006) Protein refolding in silico with atom-based statistical potentials and conformational search using a simple genetic algorithm. J Mol Biol;359(5):1456-1467.
2. Kortemme T., Morozov A.V., Baker D. (2003) An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. J Mol Biol; 326(4):1239-1259.
3. Lazaridis T., Karplus M. (1999) Effective energy function for proteins in solution. Proteins; 35(2):133-152.
4. Fang Q., Shortle D. (2005) A consistent set of statistical potentials for quantifying local side-chain and backbone interactions. Proteins; 60(1):90-96.
5. Canutescu A.A., Dunbrack R.L. (2003) Jr. Cyclic coordinate descent: A robotics algorithm for protein loop closure. Protein Sci; 12(5):963-972.
6. Lovell S.C., Word J.M., Richardson J.S., Richardson D.C. (2000) The penultimate rotamer library. Proteins ; 40(3):389-408.

Softberry - 196 models for 96 3D/100 DR targets

Softberry tools for protein structure analysis and modeling

V. Solovyev [1,2], D. Affonnikov[2], A. Bachinsky[2], I. Titov[2], N. Bakulina, V. Ivanisenko[2] and Y. Vorobjev[2]

[1]Department of Computer Science, Royal Holloway, University of London, Egham, Surrey TW20 0EX,UK; [2]Softberry Inc., 116 Radio Circle, Suite 400 Mount Kisco, NY 10549, USA,
victor@cs.rhul.ac.uk

We have developed a suite of programs that were applied to analyze CASP7 targets. These programs can be used within window based ***Molquest*** computer package or run on the web server at www.softberry.com. Identification of disordered regions in proteins was computed by the ***Pdisorder*** program that uses a combination of neural network (NN), linear discriminant function (LDF) and a smoothing procedure. At the first stage, we compute features in a sliding window of 31 residues for neural network and for the linear discriminant function At the second stage, we apply a smoothing procedure that computes chances for the positions of query sequence to be in ordered regions. The accuracy of our disorder regions predictor ***Pdisorder*** on several test sets is higher (~75-80%) than that for the other disorder fragments identification programs such as PONDR and GlobPlot.

Initial step in 3D modeling is selection of a template structure for a query sequence, or selection of a set of most similar fragments if we study a new fold, and obtaining template-query sequence alignment. This step is performed by ***Ffold*** program. ***Ffold*** alignment is made taking into account sequence similarity, secondary structures of both query and template protein, and solvent accessibility of a template protein. Secondary structure of a query protein is predicted by ***PSSFinder*** program. Secondary structure and accessibility for a template is calculated by ***SSENVID*** program. As a result, a set of aligned template-query sequence pairs is obtained. Each alignment generates a model structure, and usually up to 2-4 template-query pairs are selected for further modeling.

Building side chain and loop coordinates for a query protein based on a template structure and sequence alignment is performed by ***Getatoms*** program. To generate a set of side chain conformations for side chain structure prediction, the program uses backbone-independent rotamer library. Rotamers for each residue are ranked according to their frequency of occurrence (statistical potential) and energy of interaction with backbone (VDW scoring potential). Unfavorable conformations are then filtered out using several single-residue criteria, pairwise VDW interaction energy, and Goldstein DEE algorithm [1]. For remaining rotamers, an optimization procedure is performed to obtain a conformation with minimal VDW energy. The loop modeling procedure in ***Getatoms*** program is as follows. A large set of loop main chain conformations satisfying geometrical loop closure criteria is generated and ranked according their sterical energy of interaction with other parts of protein molecule. Top set of the conformations is subjected to the side chain optimization procedure as described above. A conformation with minimal energy is selected as loop model. This procedure is applied consequently for all the loops modeled.

Models built by ***Getatoms*** program are further refined by ***Hmod3dMM*** program, which performs energy minimization using AMBER force field (2). Hmod3dMM consists of two modules. The first module prepares a molecule topology file, which is then used as an input for molecular mechanical minimization module. Energy minimization is first performed in vacuum, and afterwards the resultant structure is further minimized in water. To handle water-water solvent interactions, Hmod3dMM employs special routines that are considerably faster than the standard ones (TIP3P/TIP4P).

In the absence of significant homology with known protein structures the structure of query protein is modeled using the ***Cover3D. Cover3D*** uses ***Ffold*** results to cover a query sequence with short similar protein fragments with known 3D structure. It outputs several variants of such coverage, which are used for manual building or computing by ***Abini3D*** a putative 3D model of target sequence. ***Abini3D*** finds optimal conformation of a set of 3D-fragments representing a target sequence. First, it generates a set of distinctive partially compact conformations, which are then optimized by genetic algorithm using

simplified model of amino acid residues. Then, the algorithm optimizes the energy function derived from statistics on known tertiary structures. Finally, ***Abini3D*** restores loop structures and outputs the atomic coordinates of optimal conformation. Resulting models are subjected for further refinement using ***Hmod3dMM*** program.

1. Goldstein R.F. (1994) Efficient rotamer elimination applied to protein side-chains and related spin glasses. Biophys J. 66, 1335-1340.
2. Weiner S.J., Kollman P.A., Nguyen D.T., Case D.A. (1986) An All Atom Force Field for Simulations of Proteins and Nucleic Acids J. Comput. Chem., 7, 230-252.

SP4 - 500 models for 100 3D targets

Template-based Protein Structure Prediction by SP[4]

Song Liu[1], Chi Zhang[1,2] and Yaoqi Zhou[1,2]

[1]HHMI Center for Single Molecule Biophysics, Department of Physiology & Biophysics, State University of New York, Buffalo, NY 14214, USA [2] Indiana University School of Informatics, Indiana University-Purdue University and Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Walker Plaza Building, Suite 319, 719 Indiana Ave., Indianapolis, IN 46202, USA

yqzhou@iupui.edu

Recognizing the structural similarity without significant sequence identity (called fold recognition) is the key for bridging the gap between the number of known protein sequences and the number of structures solved. Previously, we developed a fold-recognition method, called SP[3], which combines sequence-derived sequence profiles, secondary-structure profiles and residue-depth dependent, structure-derived sequence profiles[1]. The use of residue-depth-dependent profiles makes SP[3] one of the best automatic predictors in CASP 6. Because residue depth and solvent accessible surface area (solvent accessibility) are complementary in describing the exposure of a residue to solvent, we test whether or not incorporation of solvent-accessibility profiles into SP[3] could further increase the accuracy of fold recognition.

In this work, we address this question by developing a fold recognition method, called SP[4]. SP[4] integrates sequence-derived profiles, secondary structure profiles, residue depth-dependent structure-based profiles and solvent accessibility (SA) profiles to recognize structural homologs. Here, the solvent accessibility of query sequence with a two-state classification (buried and exposed based on a 25% SA threshold) is predicted by SABLE[2]. The residue SAs of templates are obtained by the ACCESS algorithm[3].

Table 1 shows the result of SP[4] on the Lindahl benchmark[4].

Table 1: Lindahl Benchmark (976 proteins): the summed MaxSub score for the first ranked models.

Method[a]	SP[1]	SP[2]	SP[2+]	SP[3]	SP[4]
Total	328.6	340.8	343.4	349.2	352.5
Family	286.8	292.8	292.7	293.5	295.6
Superfamily	87.5	94.3	99.9	100.8	108.9
Fold	21.7	27.8	30.2	34.5	37.1

[a]SP[1]: Sequence profiles only; SP[2]: Sequence profiles and secondary-structure profiles. SP[2+]: Sequence profiles, secondary-structure profiles, and solvent-accessibility profiles. SP[3]: sequence-derived sequence profiles, secondary-structure profiles and residue-depth dependent, structure-derived sequence profiles. SP[4]: sequence-derived sequence profiles, secondary-structure profiles, and solvent-accessibility profiles, and residue-depth dependent, structure-derived sequence profiles.

1. Zhou H, Zhou Y. (2005) Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins 58, 321--328.
2. Adamczak R., Porollo A., Meller J. (2004) Accurate prediction of solvent accessibility using neural networks-based regression. Proteins, 56, 753—767
3. Lee B., Richards F. (1971) The interpretation of protein structures: estimation of static accessibility. J. Mol. Biol. 55, 379--400.
4. Lindahl E., Elofsson A. (2000) Identification of related proteins on family, superfamily and fold level. J. Mol. Biol. 295, 613-625.

SSU - 125 models for 25 3D targets

Protein tertiary structure prediction using ECEPP/SM potential energy function and Monte Carlo with minimization

Seung-Yeon Kim[1,2], Taek-Kyun Kim[3], Julian Lee[3] and Kwang-Hwi Cho[1,3*]

[1]- Computer-Aided Molecular Design Research Center, Soongsil University, Seoul 156-743, Korea, [2]- School of General Education, ChungJu National University, Chungju 380-702, Korea, [3]- Department of Bioinformatics and Life Science, Soongsil University, Seoul 156-743, Korea,
**correspondence to: chokh@ssu.ac.kr*

For blind prediction of tertiary structures of potentially 'new fold' CASP7 targets, we have performed folding simulations based on ECEPP/SM[1] potential energy function and Monte Carlo with minimization[2]. As initial conformations of our folding simulations, we have used conformations collected from CASP7 server predictions. After folding simulations, we have obtained final conformations quite different from the initial conformations. Then we have sorted the final conformations according to their ECEPP/SM energies, and we have chosen the top-ranking conformations as our models.

ECEPP/SM potential energy function is a hybrid model based on ECEPP/3[3]. For backbone atoms except C_α, it has all atom representation, whereas, the rest of the atoms including the side chain has a reduced representation. The C_α atoms and the side chain atoms are reduced up to three pseudo atoms. C_α and the hydrogens attached to the C_α are reduced into one pseudo atoms at C_α position. C_β and the hydrogens attached to the C_β are reduced into one pseudo atoms at C_β position. For the side chain atoms beyond β position are reduced to one pseudo atoms for each amino acid. As a consequence, the model has only one χ angles for the amino acids which have more than C_γ carbons. The Hydrogen attached to C_α is contributed to only nonbonding interaction with backbone atoms within the residue in order to represent backbone torsions. The hydrogen attached to C_α and C_β carbons are contributed to only nonbonding interaction with backbone atoms within the residue in order to represent backbone torsions. The potential energy function and parameters for the backbone atoms which have all-atom representation are taken from ECEPP/3 potential energy function. The parameters for pseudo atoms were newly derived.

1. Cho K.-H., Lee B. A Simplified Potential Energy Function for ab initio Protein Folding, In preparation
2. Li Z. and Scheraga H.A. (1987) Proc. Natl. Acad. Sci., 84, 6611-6615
3. Nemethy G., Gibson K.D., Palmer K.A., Yoon C.N., Paterlini G., Zagari A., Rumsy S., and Scheraga H. A. (1992) J. Phys. Chem., 96,6472-6484

Sternberg - 190 models for 99 3D targets

Integrating *ab* initio folding, domain boundary prediction and in-house ensemble fold recognition in Phyre

L.A. Kelley[1], A. Herbert[1], R.M. Bennett-Lovsey[1] and M.J.E. Sternberg[1]

[1] – Structural Bioinformatics Group, Division of Molecular Biosciences, Imperial College London, SW7 2AY, United Kingdom
l.a.kelley@imperial.ac.uk

Our automated system is an integration of three techniques: an ensemble fold recognition system, recursive domain boundary identification, and an *ab initio* folding simulator.

Fold recognition in Phyre

Our ensemble fold recognition system uses 10 profile-profile and sequence-profile matching methods. 3D models from these systems are clustered using a novel strategy which combines measures of structural similarity between models (an 'entropic' measure) and fold recognition confidence scores (an 'enthalpic' measure), which is an adaptation of the colony-energy approach used in loop modelling[1].

Domain boundary prediction, loops and sidechains

Confident fold recognition predictions are used to define domain boundaries which are then used to split the sequence for subsequent iterations of the system. Insertions and deletions are modelled using a loop library. Loops are refined using cyclic-coordinate descent. Large loops are modelled by fragment insertion techniques. Finally, sidechains are added using the R3 algorithm[2] in conjunction with a backbone-dependent rotamer library[3].

***Ab initio* folding in NOVA**

In cases where fold recognition fails to identify a confident match and the protein sequence is <120 amino acids in length, our *ab initio* folding system (NOVA) is applied. We use fragment insertion techniques in the context of various statistical potentials. We use two *evolutionary* potentials where, instead of the target sequence, a profile of the sequence is used in assessing pair terms and solvation terms. A novel scheme to explore the register of beta-sheet structures is employed. A strand in a sheet is approximated to lie on the surface of a large circle (in order to include twisting in the sheet). The strand is

permitted to move along this circle permitting it to sample different registers on the sheet. Strand-strand packing potentials and torsion potentials are also applied. The final resulting models are clustered using a 3D-Jury approach[4].

Human predictions and automatic predictions

We have registered 2 automatic groups: Phyre-1 and Phyre-2. Phyre-1 uses a single profile-profile matching algorithm and loop modeling. Phyre-2 is the fully integrated system described above. For our manual predictions as the Sternberg group, the above Phyre-2 techniques were augmented by correcting obvious programming problems, clustering automatic predictions from other servers, and building 10,000 *ab initio* models as opposed to the 500 produced by our automatic system.

1. Xiang Z., Soto C.S. & Honig B. (2002) Evaluating conformational free energies: The colony energy and its application to the problem of loop prediction. Proc. Natl Acad Sci USA, 99(11):7432-7.
2. Xie W. & Sahinidis N.V. (2006) Residue-rotamer-reduction algorithm for the protein side-chain conformation problem, Bioinformatics, 22(2), 188-194.
3. Dunbrack Jr, R.L. & Karplus M. (1993) Backbone-dependent Rotamer Library for Proteins: Application to Side-chain prediction. J. Mol. Biol. 230, 543-574.
4. Ginalski K., Elofsson A., Fischer D. & Rychlewski L. (2003) 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics. 19(8):1015-8.

TASSER - 1027 models for 100 3D/100 QA/8 TR targets

TASSER for protein structure prediction in CASP7

H. Zhou, S.Y. Lee, S. Pandit, H. Chen, J. Borreguero, and J. Skolnick

Center for the Study of Systems Biology, School of Biology, Georgia Institute of Technology, Atlanta, USA

hzhou3@gatech.edu

The TASSER (Threading/ASSEmbly/Refinement) method (1) was further developed by using additional threading methods of SPARKS (2) and SP3 (3) as well as our previously used PROSPECTOR (4) , ab initio folded chunks for hard targets, and a new model ranking method to select models from multiple runs. The 3D-jury algorithm (5) was used for ranking the unrefined models from the three threading methods. Targets were classified as hard if the first models (ranked by Z-score from individual threading method) from all three methods have a TM-score < 0.4 with respect each other (6). In this case, we use fragment assembly method (7) to fold chunks of the target selected using an extension of the SP3 method to select significant fragment matches. Each chunk contains three consecutive segments of regular secondary (helix, strand) structure. Chunk models were ranked by comparing each position's 9 residue fragment with the corresponding fragments in the fragment library. The average RMSD was used as the ranking score. Ten models were selected for each chunk and they were used to extract consensus sequence specific contact potentials and distance restraints for TASSER to build full length models. Unlike TASSER in CASP6 which selected the top five clusters by SPICKER (8) for submission, in CASP7 we used different protocols to run TASSER multiple times and used the same fragment comparison method in the above chunk model selection for ranking all the top five full length models by SPICKER from these multiple runs. The side-chains are rebuilt using PULCHAR (9). The top five ranked models were submitted. The ranking procedure is fully automated and was also used for quality assessment prediction for all server models in CASP7.

1. Zhang Y. and Skolnick J. (2004) Automated structure prediction of weakly homologous proteins on genomic scale. Proc. Natl. Acad. Sci. (USA) 101, 7594-7599.
2. Zhou H. and Zhou Y. (2004) Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. Proteins 55, 1005-1013.
3. Zhou H. and Zhou Y. (2005) Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins 58, 321-328.
4. Skolnick J, Kihara D. and Zhang Y. (2004) Development and large scale benchmark testing of the PROSPECTOR 3.0 threading algorithm. Proteins 56, 502-518.
5. Ginalski K. and Elofsson A. and Fischer D. and Rychlewski L. (2003) 3D-jury: a simple approach to improve protein structure predictions. Bioinformatics 19, 1015-1018
6. Zhang Y. and Skolnick J. (2004) A scoring function for the automated assessment of protein structure template quality. Proteins 57, 702-710.
7. Simons K. and Kooperberg C. and Huang E. and Baker D. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 268, 209-225.
8. Zhang Y. and Skolnick J. (2004) SPICKER: a clustering approach to identify near-native protein folds. J. Comput. Chem. 25, 865-871.
9. Rotkiewicz P. and Skolnick J. Protein Chain Restoration Algorithm, in preparation.

TENETA - 225 models for 98 3D targets

TENETA - HMM-oriented Structure Prediction Method

Yuriy Sharikov[1] and Yekaterina Sharikova [2]
[1] - San Diego Supercomputer Center, [2] - San Diego State University
sharikov@sdsc.edu

In TENETA, structure prediction is obtained in several steps. During the first step, we performed BLAST-like search using primary sequences proteins, which are members of PBD. Furthermore, we used hmmpfam program from HMM package [1] for search on HMM-library. The library is built (hmmbuild –fast –gapmax 0.5) based on SCOP classification [2] and obtained alignments using Threader 3.51 program [3], a source for pairwise alignments. In the case of a considerable score (E-value ≤ 5), the search is finished and building of pdb-file occurs using MODELLER program (version 7.7 and 8.1) [4]. Input alignment always is built using TDB-files library [5], which allows disregarding of the alignment's length in resulting hmmpfam-file.

In a case of unsubstantial score, we mark the first 100 proteins from the list, which are sorted by score, with unrepeated SCOP-id (like a.4.5.x). These templates are sufficient for proper structure prediction for 95 percent of the target cases. Then, package structure building occurs [4] and we assess the derived models using evaluation program, similar to verify3D [6]. If all models are scored low, we launch an additional target search in nr (non-redundant) protein DB [7]. From the high-score protein sequences, an alignment is build, using program ClustalW [8]. Furthermore, we pair like compress obtained target-alignment, as well as, each alignment from library, used for preparing HMM. For compression, we use a variation of popular algorithm Ziv-Lempel [9]. In each case, the results for compressed target-alignment sizes and compressed alignment from the library (two compressed files' sizes) are summed up. Then, two alignments are combined; the final file is compressed as well. The difference between the sizes of compressed file and the sum of the separate file's sizes is used as a "score" (where the bigger the difference, the better).

In such matter, the best alignment is predicted from the protein alignment library with known structures; therefore, from predicted alignment, title and dominant proteins are used to build models [10]. The best model is taken as a result.

1. Eddy S.R. (1998) Profile hidden Markov models. Bioinformatics 14, 755-763
2. http://scop.berkeley.edu/
3. http://bioinf.cs.ucl.ac.uk/threader/
4. http://www.salilab.org/modeller/
5. http://bioinf.cs.ucl.ac.uk/threader/maketdbform.html
6. http://nihserver.mbi.ucla.edu/Verify_3D/
7. http://www.ncbi.nlm.nih.gov/BLAST/blastcgihelp.shtml#protein_databases
8. http://www.ebi.ac.uk/clustalw/
9. Jacob Ziv and Abraham Lempel. (1978) Compression of Individual Sequences Via Variable-Rate Coding, IEEE Transactions on Information Theory, Sep.
10. Tsigelny I., Sharikov Y., at all. (2002) HMM-based system (HMMSPECTR) for detecting structural homologies on the basis of sequential information. Protein Eng. 15, 347-352

tlbgroup - 20 models for 14 3D/ 1 FN targets

Combining homology recognition and knowledge based modelling with ensembl generation.

David F. Burke[1], Richard Bickerton[1], Alan Brown[1], Tammy Cheng[1], Nick Furnham[1], Amiram Goldblum[1,2], Sheena Gordon[1], Deepti Gupta[1], Anjum Karmali[1], Kenji Mizuguchi[1,3], Younes Mokrab[1], Rinaldo Wander Montalvao[1], Wataru Nemoto[1], Ricardo Nunez[1], James Park[1], Eva Maria Priego[1], Richard Smith[1], Duangrudee Tanramluk[1], Catherine Worth[1] and Tom Blundell[1]
1 -Crystallography and Bioinformatics group, Biochemistry Dept , University of Cambridge,80 Tennis Court Road, Cambridge CB2 1GA
2- Department of Medicinal Chemistry & Natural Products, The Hebrew University of Jerusalem, Israel
dave@cryst.bioc.cam.ac.uk

Here we combine the use of the FUGUE1 homology recognition program with two homology modelling procedures, ORCHESTRAR2 and RAPPER3.

First, fold recognition is performed using FUGUE which searches profiles derived from the HOMSTRAD4 database to produce a sequence structure alignment. Following manual inspection and adjustments, conserved structural cores of templates are defined by CHORAL5. CHORAL uses environment specific substitution tables (ESSTs) combined with differential geometry and pattern recognition algorithms to identify structurally conserved sections of superposed parent structures.

Structurally variable regions are then predicted by CODA[6] and SEARCHSLOOP[7]. CODA is a consensus approach for predicting structurally

variable regions of protein models consisting of two algorithms, FREAD and PETRA. FREAD is a knowledge-based approach that uses a fragment database consisting of all continuous thirty residue backbone segments contained in structures found in the HOMSTRAD database. PETRA is an ab initio algorithm that constructs fragments using eight phi-psi pairs, derived from the partitioning of six larger regions the Ramachandran plot. SEARCHSLOOP allows the user to search the Sloop fragment database for loop conformations connecting elements of protein secondary structure. The database consists of 80000 loops from 9000 structures from HOMSTRAD (May 2004) clustered into 3800 classes. Environmental specific sequence profiles have been calculated for every class. Each class also contains information detailing the local secondary structure environment and the angle and distances between secondary structure vectors.

Predictions of the conformations of sidechains are made by Andante. It utilizes ESST information based on observed side chain chi angle conservation from a large number of families in the HOMSTRAD database. Andante automatically restricts rotamer library solutions based upon analogous sidechains found in the parent structures.

Finally, RAPPER is used to build ensembles of models based upon the knowledge based predictions.

1. Shi J., Blundell T. L. & Mizuguchi K. (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 310, 243-57.
2. Williams M.G., Shirai H., Shi J., Nagendra H.G., Mueller J., Mizuguchi K., Miguel R.N., Lovell S.C., Innis C.A., Deane C.M., Chen L., Campillo N., Burke D.F., Blundell T.L., de Bakker P.I. (2001) Sequence-structure homology recognition by iterative alignment refinement and comparative modeling. Proteins. Suppl 5:92-7.
3. Furnham N., de Bakker P.I.W, DePristo M.A., Burke D.F., Blundell T.L. Application of RAPPER to Comparative Modelling *unpublished*
4. Stebbings L.A. & Mizuguchi K. (2004) HOMSTRAD: recent developments of the Homologous Protein Structure Alignment Database. Nucleic Acids Res 32, D203-7.2.
5. Montalvao R.W., Smith R.E., Lovell S.C. & Blundell T.L. (2005). CHORAL: a differential geometry approach to the prediction of the cores of protein structures. Bioinformatics 21, 3719-25.3.
6. Deane C.M. & Blundell T.L. (2001) CODA: a combined algorithm for predicting the structurally variable regions of protein models. Protein Sci 10, 599-612.4.
7. Burke D.F., Deane C.M. & Blundell T.L. (2000) Browsing the SLoop database of structurally classified loops connecting elements of protein secondary structure. Bioinformatics 16, 513-9.5.

Tripos-Cambridge - 13 models for 19 3D targets

ORCHESTRAR Homology Modeling

R.E Smith[1,2], Mike Dolan [2,3], Simon Cross [2,4],

[1] Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge, UK, CB2 1GA ,Tripos Inc. [2] –Tripos Inc,St.Louis

[1]res50@mole.bio.cam.ac.uk, [3]mdolan@tripos.com, [4] scross@tripos.com

Introduction.

ORCHESTRAR (commercialized by Tripos Inc.) comprises a suite of tools following the iterative process for the homology modeling of proteins, with the underlying theme of a knowledge-based approach using the information in HOMSTRAD[1]. The major components of the package include the programs; CHORAL[2], CODA[3], SEARCHSLOOP[4], ANDANTE and HARMONY3. These packages were used in conjunction with the FUGUE[5] homology recognition program. The user is provided with an ensemble of structurally conserved regions extracted from superposed parent structures. Structurally variable regions are then modeled by any one of three programs that access different loop solutions. Side-chain placement is aided by the use of parent information. Sequence-structure alignment evaluation and model validation is then performed. Poorly modeled regions are then reassessed.

Methodology.

1. Homology Recognition

Performed by the program FUGUE.

2. Core construction

CHORAL uses a knowledge-based method consisting of differential geometry and pattern recognition algorithms to identify structurally conserved sections of superposed parent structures. Environment specific substitution tables (ESSTs) are used to classify and filter which patterns likely to represent the core target. The environments for the substitution tables are defined for the backbone geometry of the parents.

3. Loop building

CODA is a consensus approach for predicting structurally variable regions of protein models. The two algorithms, FREAD and PETRA, are used to predict loop solutions. FREAD is a knowledge-based approach that uses a fragment database consisting of all continuous thirty residue backbone segments contained in structures found in the HOMSTRAD[1] database. Selection filters include; Cα separation of anchor residues, anchor residue rmsd, energy term for the superposed fragment and an environmentally constrained substitution score.

Six phi-psi regions of the Ramachandran plot define the environments considered. PETRA is an algorithm that constructs loop solutions ab initio. Loop solutions consist of fragments constructed from eight phi-psi pairs, giving a maximum of $(n+4)^8$ possible loops for any gap of n residues (+ 4 anchor residues). The determination of these phi-psi pairs resulted from the calculation of individual amino acid propensities for partitions of six larger regions the Ramachandran plot. The CODA method then does a pair wise comparison of all FREAD and PETRA predictions. For consensus results a loop pair must pass a number of filters including difference of backbone torsion angles and sum of energy in superposed position.

SEARCHSLOOP allows the user to search the Sloop fragments database for loop conformations connecting elements of protein secondary structure. The database consists of ~80000 loops from ~9000 structures from HOMSTRAD (May 2004) clustered into ~3800 classes. Each class contains information about its member loops, such as local secondary structure environment and the angle and distance between secondary structure vectors. Scoring is based on anchor rmsd and environment specific substitution scores.

4. Side Chain Placement

This is performed by the program ANDANTE. It utilizes ESST information based on observed side chain chi angle conservation of a large number of families in the HOMSTRAD database. Depending on the parent-target residue substitution, this information allows Andante to borrow entire high probability side-chain conformations or to restrict rotamer library solutions to specific chi bins. Side chain placement for non-borrowed positions is done by an interacting cluster/simulated annealing approach.

5. Model validation (Error detection in sequence-structure alignment)

HARMONY3 is used to locate errors that may have occurred in the sequence-structure alignment that have been carried through the model building process. The Harmony3 score for each modelled residue is calculated and consists of five components; the amino acid propensity score for the observed environment, observed amino acid distribution for that residue obtained from a PSI-BLAST search, a propensity score for a residue based on observed ESST scores. A composite substitution score from merged ESSTs. Finally, a term for local alignment flexibility is calculated. This takes into account the number of gaps in a window around a position in the sequence-structure alignment and also the number of identical residue pairs in the window is incorporated into the score. Low scoring regions are then re-examined for errors in the sequence-structure alignment or modeling errors.

1. Stebbings L.A. & Mizuguchi K. (2004) HOMSTRAD: recent developments of the Homologous Protein Structure Alignment Database. Nucleic Acids Res 32, D203-7.
2. Montalvao R.W., Smith R.E., Lovell S.C. & Blundell T.L. (2005) CHORAL: a differential geometry approach to the prediction of the cores of protein structures. Bioinformatics 21, 3719-25.
3. Deane C.M. & Blundell T.L. (2001) CODA: a combined algorithm for predicting the structurally variable regions of protein models. Protein Sci 10, 599-612.
4. Burke D.F., Deane C.M. & Blundell T.L. (2000) Browsing the SLoop database of structurally classified loops connecting elements of protein secondary structure. Bioinformatics 16, 513-9.
5. Shi J., Blundell T. L. & Mizuguchi K. (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 310, 243-57.

Tsailab - 245 models for 42 3D/7 TR targets

Measuring 3D information from protein structure: "3D-bits", an intuitive yet quantitative assessment of comparative modeling predictions

Rosemarie Swanson[1], Jerry Tsai[1]

[1]-Texas A&M University

rosmar@tamu.edu

Comparative modeling is an attempt to predict an unknown 3-d structure associated with an amino acid sequence by adjusting a known 3-d structure of a protein with a similar amino acid sequence. However, it has not been clear that such an adjusted structure improves on the unadjusted known structure. In this work we present a method of measuring the information that one 3d protein structure supplies about another structure and show that by this measure, comparative modeling predictions show a statistically significant improvement from CASP4 to CASP6. By this new measure, for the targets we examined, in CASP6 about 80% of the assessor-top-scored models described the experimentally determined target structure better than the best unadjusted parent structure did, while in CASP4 only half the best models were better than the best parent.

A characteristic of 3-d superpositions of molecules is that generally the more atoms that are superimposed the less precise the match becomes. There is a trade-off. The trade-off problem has been addressed by GDT_TS, which reports the average of the percentages of atoms superimposed within four different cutoff distances. A small protein has a built-in advantage over a large one by this measure simply because fewer atoms have to be made to match.

Starting from a different point of view, we arrived at a structure comparison measure that is complementary to GDT_TS, but is finer-grained, and reflects

(1) the greater difficulty of predicting a larger structure, and (2) the greater difficulty of improving on an already good match.

The basic idea is that the probability of a predicted atom falling close to its target atom by chance is related to the volume around the target atom within which it falls (so that one should use the cube of atom mismatch distance (rmsd) instead of the linear rmsd to measure how good the match between atoms is). We used the ratio of the mismatch volume to the total protein volume as the probability of a chance match, and summed the logarithms of these probabilities over the predicted atoms to obtain a 3D-bit score for the goodness of match between two structures. The mismatch volume for each pair of atoms was determined by finding the best superposition between the two structures,(withA Zemla's LGA program), then using the distance between each pair of atoms as the radius of a sphere that measures the mis-match volume for that pair.

The "3D-bit" approach is similar to the GDT_TS approach in that each gives a higher weight to contributions from closer matches, and each score is improved by having more matching atoms.

The 3D-bit measure differs from the GDT_TS measure in two significant ways. First, smaller mismatches contribute relatively more to the 3D-bit score than they do to the GDT_TS score. Cutting a mismatch distance in half reduces its mismatch volume by a factor of 2 to the third power, whereas the GDT_TS score improves at most linearly. Furthermore, for matches closer than 1 Angstrom (for example), further reducing the mismatch distance contributes no improvement to the GDT_TS score because the GDT_TS score is threshold-based. So the 3D-bit score is more sensitive to small improvements in matching. Furthermore, the 3D-bit score acknowledges that the quarter-Angstrom improvement in rmsd from 0.75 A to 0.50 A represents a greater improvement in prediction skill than the quarter-Angstrom improvement from 1.0 A to 0.75 A. Secondly, the 3D-bit score reflects the absolute number of atoms matched, and it increases for a larger number of predicted atoms, so that it reflects the intrinsic difficulty of the prediction problem and can be used to compare the amount of information provided by predictions even for different targets. Since GDT_TS is expressed as a percent, it doesn't express the greater difficulty of predicting larger targets.

We describe "3D-bits" and the results of applying it to a collection of pairs of structures, which included a motley collection of 27 globin chains, and all possible comparisons between the experimentally-determined target structure, the top-scoring model, and the parent structures for 33 single-domain comparative modeling targets -- 7 from CASP4 and 26 from CASP6.

Tsailab - 245 models for 42 3D/7 TR targets

Side-Chain Guided Protein Refinement

Xiaotao Qu, Rosemarie Swanson, Zach Bohannan, Robert Bliss, and Jerry Tsai

Texas A&M University, Center for Structural Biology, Dept. of Biochemistry & Biophysics

xiaotao@tamu.edu, rosmar@tamu.edu, zee_bo@tamu.edu, standbob@gmail.com,jerrytsai@tamu.edu

In this years CASP 7 experiment, our group (the Tsailab #1273-7338-1989) focused on comparative modeling and made submissions for 43 targets. Unlike backbone directed methods, our approach considers the variation in side-chain packing, and uses this variation to direct the moves in the refinement of protein structure[1]. Using a Voronoi polyhedra approach[2,3], we identify interacting residues within a protein. By treating each residue as a node and an interaction as an edge, we generalize the protein core into a graph and thereby can group residues into cliques, where a clique is a set of that all interact with each other. This clique approach allows us to characterize the minimal packing unit as residues that all contact each other. Comparing these cliques between protein structures defines a relative packing group. These relative packing groups were found for two sets of structures: 1) within homologous structures to the target sequence and 2) across the PDB[4]. The relative packing groups from homologous set were selected only if they existed in 50% or more structures. The relative packing groups from the PDB were clustered based on secondary structure of the residues and a CαRMSD cutoff of 0.5 Å. From these relative packing groups, we calculated non-local distance constraints as inputs into a distance geometry algorithm to create target structures. We used relative packing groups from homologous structures exclusively except in cases where the non-local constraints are underdetermined (fewer than 5 homologous structures). In such cases, matching relative packing groups from the PDB data set were selected to help augment the homologous data set. We implemented these non-local constraints in a structure refinement algorithm in the following procedure. The MUSTANG algorithm aligned structures[5], and FASTA aligned the target sequence to the sequences[6] from the homologous structures. These alignments mapped the non-local constraints generated from the relative packing group as well as a set of local and torsion constraints analysis onto the target structure. An initial structure was generated using MODELLER[7]. The starting structure and constraints were used as inputs to the distance geometry/simulated annealing routine in XPLOR-NIH[8] and on average ~250 structures were generated. Candidate structures are clustered based on CαRMSD similarity to each other and the largest five clusters are chosen for further scoring. Within each of these five clusters, two scores are used to evaluate the candidate structures. Based on molecular dynamics simulations of

a protein fold set (100 ns total simulation of 125 protein folds), we compiled probability distributions of side-chain volume based on backbone torsion angles and the propensity of a residue's χ_1 angle also based on backbone torsion angle. The structures with the top 15-20 best scores were then viewed by eye and 5 structures were chosen for submission.

1. Holmes J.B. & Tsai J. (2005) Characterizing Conserved Structural Contacts by Pair-wise Relative Contacts and Relative Packing Groups. J Mol Biol 354, 706-21.
2. Harpaz Y., Gerstein M. & Chothia C. (1994) Volume changes on protein folding. Structure 2, 641-9.
3. Voronoi G.F. (1908) Nouveles applications des paramétres continus à la théorie des formes quadratiques. J. Reine Angew. Math. 134, 198-287.
4. Berman H.M. et al. (2000) The Protein Data Bank. Nucleic Acids Res 28, 235-42.
5. Konagurthu A.S., Whisstock J.C., Stuckey P.J. & Lesk A.M. (2006) MUSTANG: a multiple structural alignment algorithm. Proteins 64, 559-74.
6. Pearson W.R. (2000) Flexible sequence similarity searching with the FASTA3 program package. Methods Mol Biol 132, 185-219.
7. Fiser A. & Sali A. (2003) Modeller: generation and refinement of homology-based protein structure models. Methods Enzymol 374, 461-91.
8. Schwieters C.D., Kuszewski J.J., Tjandra N. & Clore G.M. (2003) The Xplor-NIH NMR molecular structure determination package. J Magn Reson 160, 65-73.

UCB-SHI - 885 models for 96 3D/98 QA targets

Modeling and Quality Assessment using HARMONY3 and QUAD

Jiye Shi[1]

[1]*-UCB Inc.*

jiye.shi@gmail.com

Homology Modeling – FUGUE and other fold recognition tools were used to identify potential templates and to generate initial alignments. MODELLER was used for model building. HARMONY3 (unpublished), an algorithm designed to detect problematic alignment regions, was used to select the best template and the best alignment. QUAD (unpublished; see Quality Assessment below) was then used to evaluate the models and to select the most promising models. Human intervention was limited to no more than 2 hours per target, which was mainly directed to improving the final alignment chosen by HARMONY3.

New Fold – Potential structural fragments were identified using multiple algorithms such as fold recognition and ab initio modeling. The fragments were then manually assembled and the resulting models were evaluated by QUAD (unpublished; see Quality Assessment below). Models with highest QUAD scores were selected and submitted. Human intervention was limited to no more than 2 hours per target.

Quality Assessment – QUAD propensity-based score describes the "fitness" of each residue to its structural environment, which is defined by secondary structure element, H-bond to backbone NH, H-bond to backbone CO and solvent accessibility. This procedure is fully automated; human intervention was limited to corrections of formatting errors and results submission. Limitations: this method works best on models with complete backbone and side chains; Large number of missing residues in AL based 3D models will result in an inaccurate quality score; CA-only models cannot be evaluated using this method.

UNI-EID_bnmx - 462 models for 100 3D targets
UNI-EID_expm - 100 models for 100 3D targets
UNI-EID_sfst – 454 models for 100 3D targets

A probabilistic approach to remote homology detection and 3D protein structure modeling

A. Poleksic[1], J.F. Danzer[2], B. Palmer[2], M.Fienup[1] and D.A. Debe[2]

[1] *– Department of Computer Science, University of Northern Iowa*

[2] *– Eidogen-Sertanty, Inc., San Diego, California*

poleksic@cs.uni.edu

UNI-EID algorithms are profile-profile methods that utilize information contained in multiple sequence alignments corresponding to the query and template's protein family. An internally modified version of PSI-BLAST[1] is used to construct sequential profiles corresponding to query sequence and each of the template sequences[2]. Each pair of profiles is then scored and aligned using a novel dynamic programming and a probabilistic scoring scheme that has an analogy in an experiment of throwing an irregular 20-sided die. Our new probabilistic alignment scoring function, tested in CASP7, also takes into account template structural information as well as predicted local structure of the query protein. The gap penalties are position specific and reflect the similarity of profiles being aligned, the aligned residues' secondary structure

states, and the distribution of gaps in PSI-BLAST multiple alignments. Statistical significance of alignment scores is computed independently for each pair of sequences using Convergent Island Statistics (CIS)[3]. The CIS method estimates score statistics "on the fly" and can be readily applied to any alignment algorithm whose background scores follow an extreme value distribution. The method contains no parameters to optimize and there is no need for fitting the data of any kind.

UNI-EID_sfst reports the best five local alignments to PDB templates. UNI-EID_bnmx uses different background probabilities when scoring pairs of profiles. This change to the background model results in slightly longer models compared to those generated by UNI-EID_sfst. In CASP7, UNI-EID_bnmx reports backbone atom coordinates derived from the single template corresponding to the highest scoring alignment. UNI-EID_expm builds unrefined protein structures from multiple PDB templates corresponding to the top scoring UNI-EID_sfst's models. The remaining backbone atoms are reconstructed from the α-carbon coordinates[4].

1. Altschul S.F., Madden T.L., Schäffer A.A., Zhang J., Zhang Z., Miller W. and Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res., 25, 3389-3.
2. Debe D.A., Danzer J.F., Goddard W.A., Poleksic A. (2006) STRUCTFAST: Protein sequence remote homology detection and alignment using novel dynamic programming and profile-profile scoring, Proteins, 64, 960-967.
3. Poleksic A., Danzer J.F., Hambly K., Debe D.A. (2005) Convergent Island Statistics: a fast method for determining local alignment score significance. Bioinformatics, 21, 2827-2831.
4. Rey A., Skolnick J. (1992) Eficient algorithm for the reconstruction of a protein backbone from the α-carbon Coordinates. J. Comput. Chem. 13, 443-456.

UWscore - 97 models for 97 QA targets

Protein Structure Quality Prediction with Support Vector Regression

Jian Qiu[1], Will Sheffler[1], David Baker[1,2], William Stafford Noble[1,3]

[1] –Department of Genome Sciences, [2] – Department of Biochemistry, [3]–Department of Computer Science and Engineering, University of Washington

jianq@u.washington.edu, wsheffle@u.washington.edu, dabaker@u.washington.edu, noble@gs.washington.edu

We applied the support vector regression (SVR) machinery to the problem of protein structure quality prediction based on 34 features generated by Rosetta[1] and T32S3, a distance-dependent atomically detailed potential[2]. The Rosetta features include measures that describe the overall shape and burial, packing, solvation effects, hydrogen bonding patterns, attractive and repulsive Van der Waals forces, and so on.

We developed the training set from CASP5 and CASP6 predictions and a non-redundant data set of native PDB structures. A pre-compiled CulledPDB list from PISCES[3] was downloaded from http://dunbrack.fccc.edu/Guoli/pisces_download.php.

A filtered subset of native structures in this list was combined with CASP5 and CASP6 structures to come up with the training set. Rosetta energy local minimization was first performed on each structure to remove clashes and optimize the rotamers of side chains. The Rosetta features and T32S3 were then computed based on these minimized structures. A structure was included in the training set only if the minimized structure shares significant structural similarity with the original structure. We trained the SVR to predict RMSD from the features of a structure. The RMSDs for the predicted structures were extracted from the CASP5 and CASP6 evaluation results. All the native structures were assigned an RMSD of 0.

We used the epsilon SVR in LIBSVM[4] as the SVR implementation, with a radial basis function (RBF) kernel of the form $K(x,y)=\exp\{-\gamma\|x-y\|^2\}$. Before predicting the quality of a structure, the structure was first subjected to Rosetta energy local minimization. Features were computed from the minimized structures. The SVR model learned from the training set was then used to predict the quality of the structure from the features based on its minimized structure.

1. Rohl C.A., Strauss C.E., Misura K.M. and Baker D. (2004) Protein structure prediction using Rosetta. Methods Enzymol. 383, 66-93.

2. Qiu J. and Elber R. (2005) Atomically detailed potentials to recognize native and approximate protein structures. Proteins. 61, 44-55.
3. Wang, G. and Dunbrack R.L. Jr. (2005) PISCES: recent improvements to a PDB sequence culling server. Nucleic Acids Research. 33, W94-98.
4. Chang C.C. and Lin C.J. (2001) LIBSVM: a library for support vector machines, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Wymore- 104 models for 37 3D targets

Comparative modeling using a stochastic alignment algorithm and statistical potentials

Adam C. Marko , Adam Kraut, Troy Wymore
Pittsburgh Supercomputing Center
wymore@psc.edu

From our past experience in the CASP6 experiment, we learned that profile-sequence bioinformatics search methods (MEME/MAST) failed to identify several template structures and that our alignment sampling (200-1000) was not extensive enough. In addition, we overemphasized the ability of molecular mechanics force fields to identify the most native structure in an ensemble. During CASP7, templates were selected from various fold-recognition servers though we did not use their alignments or structures in any manner. We then constructed 5000 pairwise alignments per target/tempate pair considering up to 4 different templates per target. The alignments were created by stochastic backtracking based on match probabilities using the program probA[1]. Three-dimensional models were then created with the program MODELLER[2]. The resultant structures were then scored with the statistical potentials DFIRE[6] and DOPE (available in the MODELLER program). The top twenty models ranked by DFIRE and DOPE from each template were then visually examined. The lowest energy models that did not have any obvious structural errors (knots, etc) were then selected for submission. In addition, the lowest 1000 energy structures as ranked by DFIRE from the most favorable template were analyzed by Prosa2003[3]. The lowest energy model as ranked by Prosa was then submitted as well.

1. Muckstein U., Hofacker I.L., Stadler P.F. (2002) Stochastic pairwise alignments. Bioinformatics, 18, S153-S160.
2. Sali A., Blundell T.L. (1993) Comparative Protein Modeling by Satisfaction of Spatial Restraints. J. Mol. Biol., 234,779-815.
3. Sippl M.J. (1993) Recognition of Errors in Three-Dimensional Structures of Proteins. PROTEINS: Struct. Func. Gen. 17,355-362.
4. Feig M., Karanicolas J., Brooks III, C.L.B. (2004) MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology. J. Mol. Graph. Model. 22, 377-395.
5. Skolnick J., Kolinski A., Ortiz A.R. (1997) MONSSTER: a method for folding globular proteins with a small number of distance restraints. J. Mol. Biol. 265, 217-241.
6. Zhang C., Liu S., Zhu Q., Zhou Y. (2005) A knowledge-based energy function for protein-ligand, protein-protein, and protein-DNA complexes. J. Med. Chem. 48(7):2325-35.

YASARA - 68 models for 21 3D/21 QA/5 TR targets

High resolution refinement with YASARA

Elmar Krieger
CMBI, Center for Molecular and Biomolecular Informatics,
Radboud University Nijmegen, the Netherlands
Elmar.Krieger@yasara.org, www.YASARA.org

The **last mile of the protein folding problem** has been approached with a number of new developments during CASP7, implemented in the framework of the molecular modeling program **www.YASARA.org**. The entire procedure was fully automatic but too time consuming to participate as a server.

First the server predictions were downloaded from the CASP site, missing loops where added with YASARA, missing side-chains rebuilt with SCWRL[1]. Then the models were ranked using the newly developed **Twinset Score**, a combination of force field and solvation energies as well as knowledge based potential energies assigned by YASARA[2,3], and the classic WHAT_CHECK[4] scores reported by WHAT IF[5]. This ranking was submitted as a quality prediction, where 1.0 corresponded to a perfect protein and 0.0 to garbage.

Then the top-scoring model was picked and subjected to thousands of parallel molecular dynamics simulations using the Models@Home distributed computing system[6] and the newly developed **YASARA force field**, a third-generation self-parameterizing energy function[2] obtained in crystal space[7] from the YAMBER force field[3]. To speed up the sampling of conformational space, some of the simulations were accelerated with CONCOORD[8]. Those models that were likely to have moved closer to the native structure during the simulation were identified with the **Twinset Cluster Score**, which employs clustering methods to remove false positives. The best model was subjected to another round of refinement until the procedure converged.

Many thanks to all supporting users of the YASARA molecular modeling program for financing this work.

1. Canutescu A.A., Shelenkov A.A. & Dunbrack R.L.J. (2003) A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci. 12, 2001-2014.
2. Krieger E., Koraimann G. & Vriend G. (2002) Increasing the precision of comparative models with YASARA NOVA - a self-parameterizing force field. Proteins 47, 393-402.
3. Krieger E., Darden T., Nabuurs S.B., Finkelstein A. & Vriend G. (2004) Making optimal use of empirical energy functions: force field parameterization in crystal space. Proteins 57, 678-683.
4. Hooft R.W.W., Vriend G., Sander C. & Abola E.E. (1996) Errors in protein structures. Nature 381, 272-272.
5. Vriend G. (1990) WHAT IF - A molecular modeling and drug design program. J.Mol.Graph. 8, 52-56.
6. KriegerE. & Vriend G. (2002) Models@Home: distributed computing in bioinformatics using a screensaver based approach. Bioinformatics 18, 315-318.
7. Krieger E., Nielsen J.E., Spronk C.A.E.M. & Vriend G. (2006) Protein pKa prediction by Ewald summation. J Mol Graph Model.
8. de Groot B. L. et al. (1997) Prediction of protein conformational freedom from distance constraints. Proteins 29, 240-251.

YU-BA - 72 models for 72 FN targets

Yuba (Yet Unnamed Binding site Assessment) aiming at the binding site prediction in protein function category of CASP7

Daisuke Takaya, Genki Terashi, Mayuko Takeda-Shitaka, Kazuhiko Kanou, Mitsuo Iwadate, Akio Hosoi, Kazuhiro Ohta and Hideaki Umeyama

Department of Biomolecular Design, School of Pharmacy,Kitasato University
p99150@st.pharm.kitasato-u.ac.jp

In the area of binding site prediction, we want to know how useful of using known protein structure having ligand. So in protein function (FN) prediction category of CASP7, Yuba team aims at the prediction of the binding site of the target protein using known structure in PDB.

Step 1: selecting "base model"

First, select "base model". All the server models were obtained from CASP7 home page (http://www2.predictioncenter.org/index_serv.html) for this purpose. These models include tertiary structure (TS) and alignment (AL). These were refined or changed to tertiary structure by FAMS[1]. If it was AL format, a model was built based on this alignment. If it was TS format, a model was refined by FAMS. We used all the server models as its template because these models include CA model or lacking residues. These refined server models were evaluated using specialized CIRCLE[2] 3D1D Score for CM, the category of which is determined from the SVM program and select the highest score model as "base model".

Step 2: superimposing

Second, obtain known PDB structure having ligand and superimpose to "base model" using CE program[3]. The list of superimposed PDB is gotten from PARENT of server. PDB not having ligand is ignored.

Step 3: clustering

After ligand atoms were extracted from superimposed structures, clustering ligand atoms by nearest neighbor method and choosing largest cluster were excuted. Atom type is ignored.

Step 4: evaluating

In this selected cluster, count the number of atoms in collision with "base model" on the condition that distance regarded as collision are 4, 6, 8 and 10Å, respectively. And rank it by the number of collision atoms in ascending sort. The size of ligand was not considered, and we chose 10 residues in listed amino acid residues of the "based model".

A part of result

Experimental determined structure list that Yuba method could submit is shown as follows. T0283, T0284, T0286, T0288, T0290, T0291, T0292, T0293, T0295, T0301, T0303, T0304, T0305,T0306,T0307,T0308, T0310, T0311, T0312, T0313, T0315, T0317, T0318, T0319, T0320, T0322, T0323, T0324, T0325, T0326, T0327, T0329, T0330, T0331, T0332, T0334, T0335, T0338, T0339, T0340, T0341, T0343, T0344, T0346, T0347, T0350, T0351, T0352, T0353, T0355, T0356, T0357, T0358, T0359, T0360, T0362, T0364, T0365, T0366, T0367, T0368, T0370, T0371, T0374, T0375, T0376, T0378, T0380, T0382, T0384, T0385 and T0386 (total 72). In target T0292, our prediction of binding site is 12, 13, 15, 16, 20, 90, 91, 143, 144, and 146. In the experimentally determined structures (PDBID: 2CL1), residues of binding site are assumed if residue is within 5.0Å of the ligand (5-[(z)-(5-chloro-2-oxo-1,2-dihydro-3h-indol-3-ylidene)methyl]-n-(diethylamino)ethyl)-2,4-Dimethyl-1h-pyrrole-3-carboxamide in 2CL1). In this condition, residues of binding site are 12, 20, 33, 35, 66, 84, 85, 86, 87, 88, 90, 91, 94, 146, 160 and 164. The residue number of correct prediction is 12, 20, 90, 91 and 146.

Discussion

Unfortunately, there are some incorrect predictions of residue in target T0292. This method is mostly depending on how to choose "base model". So the method may be not useful for FR, NF category which is difficult to predict reliable model.

1. Ogata K. and Umeyama H. (2000) An automatic homology modeling method consisting of database searches and simulated annealing. J. Mol. Graphics Mod. 18 258-272.
2. See "CIRCLE: Full automated homology-modeling server using the 3D1D scoring functions" item in this book
3. Shindyalov I.N., Bourne P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Engineering 11(9) 739-747.

Zhang - 500 models for 100 3D targets

Protein structure prediction by iterative TASSER simulations

Yang Zhang

Center for Bioinformatics and Department of Molecular Bioscience, University of Kansas, 2030 Becker Dr, Lawrence, KS 66047

yzhang@ku.edu

The human-expert prediction of our group has used a similar iterative TASSER (called I-TASSER[1]) approach as what we used in the Server Section. The main gains of the human prediction in comparison with our server prediction are: (1) We have a better domain assignment which is based on our visual view of the threading alignments combined with the domain predictions of CASP7 servers; (2) we can make use of threading results from CASP7 servers which give TASSER a more diverse set of starting conformations in comparison with only using our in-house PPA threading templates; (3) we can run TASSER simulations in a longer CPU time which allows a more extensive conformation search. The I-TASSER protocol consists of three consecutive steps.

Collection of threading templates. The collection of threading templates is the first step of I-TASSER protocol, which provide basic building blocks (continuous structure fragments) for TASSER structure reassembly as well as resources to extract spatial restraints to guide TASSER simulations.[2] Threading templates in our human prediction come from two resources: (1) Four in-house profile-profile alignment methods which have their confidence parameters pre-trained in benchmarks; (2) threading results from CASP7 servers including FUGUE,[3] HHpred,[4] mGenThreader,[5] and SP3.[6] A set of 20/30/50 templates are taken from the top hits of the servers for Easy/Medium/Hard targets. The target category is assigned based on the Z-score of four in-house PPA threading methods.

TASSER structure assembly. Based on the threading alignments, target sequences are divided into aligned and unaligned regions. The models of aligned regions are directly excised from the template proteins and allowed to rotate and translate in an off-lattice system. The unaligned regions are modeled by the TASSER ab initio component,[7] which serve as linkage points of the rigid-body rotations. The potential is similar as original TASSER,[2] which consists of predicted secondary structure from a combination of PSIPRED [8] and SAM,[9] backbone hydrogen bonds,[10] a verity of statistical short- and long-range correlations,[7] and consensus contact/distance restraints extracted from the threading alignments. The major new potential added in I-TASSER is the incorporation of predicted accessible surface area through neural network.[11] All weighting parameters of I-TASSER force field have been separately tuned in Easy/Medium/Hard categories on the basis of structural decoys.[7] The Monte Carlo trajectories generated in low temperature replicas are clustered by SPICKER. [12] The cluster centroids of the highest structural density are returned for the further I-TASSER refinement.

TASSER iteration. Following the SPICKER clustering, we run TASSER Monte Carlo simulations again, starting from the selected cluster centroids. The distance and contact restraints in the second round TASSER are pooled from the initial high-confident restraints from threading, and the restraints taken from the cluster centroid structures and the PDB structures searched by the structural alignment program TM-align [13] based on the cluster centroids. The conformations with the lowest energy in the second round are selected. Finally, Pulchra [14] is used to add backbone atoms (N, C, O) and Scwrl_3.0 [15] to build side-chain rotamers.

For multiple domain proteins, we assign the domain borders mainly based on our visual view of consensus threading alignments, which may be further adjusted by the comparison with domain server predictions by Robetta-Ginzu [16] and Ma-OPUS-DOM. I-TASSER simulations will be done for the full chain and the separate domains. The final full-length models are generated by docking the domain models together under the guide of the full-chain model of I-TASSER. The domain docking is performed by a quick Metropolis Monte Carlo simulation where the energy is defined as the RMSD of domain models from the full-chain model plus the reciprocal of the number of steric clashes between domains.

1. Wu S. T. & Zhang Y. (2006) Ab initio modeling of small proteins by iterative TASSER simulations. Submitted.

2. Zhang Y. & Skolnick J. (2004) Automated structure prediction of weakly homologous proteins on a genomic scale. Proc Natl Acad Sci U S A 101, 7594-7599.
3. Shi J., Blundell T.L. & Mizuguchi K. (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 310, 243-57.
4. Soding J. (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics 21, 951-60.
5. McGuffin L.J. & Jones D.T. (2003) Improvement of the GenTHREADER method for genomic fold recognition. Bioinformatics 19, 874-81.
6. Zhou H. & Zhou Y. (2005) Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins 58, 321-8.
7. Zhang Y., Kolinski A. & Skolnick J. (2003) TOUCHSTONE II: A new approach to ab initio protein structure prediction. Biophys J 85, 1145-1164.
8. Jones D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292, 195-202.
9. Karplus K., Barrett C. & Hughey R. (1998) Hidden Markov models for detecting remote protein homologies. Bioinformatics 14, 846-856.
10. Zhang Y., Hubner I., Arakaki A., Shakhnovich E. & Skolnick J. (2006). On the origin and completeness of highly likely single domain protein structures Proc Natl Acad Sci U S A 103, 2605-10.
11. Chen H. & Zhou H.X. (2005). Prediction of solvent accessibility and sites of deleterious mutations from protein sequence. Nucleic Acids Res 33, 3193-9.
12. Zhang Y. & Skolnick J. (2004) SPICKER: A clustering approach to identify near-native protein folds. J Comput Chem 25, 865-71.
13. Zhang Y. & Skolnick J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 33, 2302-2309.
14. Feig M., Rotkiewicz P., Kolinski A., Skolnick J. & Brooks C.L., 3rd. (2000) Accurate reconstruction of all-atom protein representations from side-chain-based low-resolution models. Proteins 41, 86-97.
15. Canutescu A.A., Shelenkov A.A. & Dunbrack, R.L., Jr. (2003). A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci 12, 2001-14.
16. Kim D. E., Chivian D., Malmstrom L. & Baker D. (2005) Automated prediction of domain boundaries in CASP6 targets using Ginzu and RosettaDOM. Proteins 61 Suppl 7, 193-200.

Zhang-Server - 500 models for 100 3D targets

Server prediction by iterative TASSER simulations in CASP7

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

All CASP7 targets are modeled by an automated tool of iterative TASSER simulation, called I-TASSER,[1] which includes three consecutive steps.

Threading. The target sequences are threaded through a non-redundant PDB structure library with the purpose of identifying appropriate global-structure templates (for CM/FR targets) or local fragments (for NF targets). Threading is done by four simple profile-profile alignment (PPA) runs, where the alignment score consists of sequence profile and secondary structure matches.[1] In the first PPA run, both profiles of target and template sequences are generated by PSI-Blast search;[2] In the second alignment run, the profiles are generated by hidden Markov model from SAM-T99.[3] During the construction of profiles, Henikoff weights are used for re-weighting the redundant sequences.[4] The Needleman-Wunsch global dynamic programming alignment algorithm [5] is used to find the best match between query and template sequences. The third and the forth PPA alignments are similar as that in the first and the second runs but the Smith-Waterman local alignment algorithm [6] is exploited.

TASSER structure assembly. 20/30/50 templates are selected from the four sets of PPA threading alignments for Easy/Medium/Hard targets, which are used for further TASSER Monte Carlo reassembly.[7] The category of Easy/Medium/Hard is assigned based on the PPA Z-scores and pre-trained on the benchmark TASSER simulations.[1] Based on the threading alignments, target sequences are divided into aligned and unaligned regions. The models of aligned regions are directly excised from the template proteins and allowed to rotate and translate in an off-lattice system. The unaligned regions are modeled by the TASSER ab initio modeling,[8] which serve as linkage points of the rigid-body movement of aligned regions. The Monte Carlo search is implemented by the parallel exchange method,[9; 10] with each replica starting from different templates. The potential is similar as original TASSER,[7] which includes predicted secondary structure from a combination of PSIPRED [11] and SAM,[3] backbone hydrogen bonds,[12] a verity of statistical short- and long-range correlations,[8] and consensus contact/distance restraints extracted from the PPA alignments. The major new potential is the incorporation of predicted accessible surface area through neural network.[13] All weighting parameters of I-TASSER force field have been separately re-tuned in Easy/Medium/Hard categories on the basis of structural decoys.[8] After TASSER simulations, the structure decoys generated in low temperature replicas are clustered by

SPICKER.[14] The cluster centroids of the highest structure density are returned for the further I-TASSER refinement.

TASSER iteration. Following the SPICKER clustering, we run TASSER Monte Carlo simulations again, which starts from the selected cluster centroids. The distance and contact restraints in the second round TASSER are pooled from the initial high-confident restraints from threading and the restraints taken from the cluster centroid structures and the PDB structures searched by the structure alignment program TM-align[15] based on the cluster centroids. The conformations with the lowest energy in the second round are selected. Finally, Pulchra [16] is used to add backbone atoms (N, C, O) and Scwrl_3.0 [17] to build side-chain rotamers.

Multiple-domain proteins. If any region with >80 residues has no aligned residues in at least two strong PPA hits, the target will be judged as a multiple-domain protein and domain boundaries are automatically assigned based on the borders of the large gaps. I-TASSER simulations will be run for the full chain as well as the separate domains. The final full-length models are generated by docking the domain models together. The domain docking is performed by a quick Metropolis Monte Carlo simulation where the energy is defined as the RMSD of domain models from the full-chain model plus the reciprocal of the number of steric clashes between domains. The goal is to find the domain docking orientation that is closest to the I-TASSER full-chain model and has the minimum steric clashes. The final models docked from I-TASSER domains are submitted to CASP7.

1. Wu S.T. & Zhang Y. (2006) Ab initio modeling of small proteins by iterative TASSER simulations. Submitted.
2. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W. & Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25, 3389-402.
3. Karplus K., Barrett C. & Hughey R. (1998) Hidden Markov models for detecting remote protein homologies. Bioinformatics 14, 846-856.
4. Henikoff S. & Henikoff J.G. (1994) Position-based sequence weights. J Mol Biol 243, 574-8.
5. Needleman S.B. & Wunsch C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48, 443-53.
6. Smith T.F. & Waterman M.S. (1981) Identification of common molecular subsequences. J Mol Biol 147, 195-7.
7. Zhang Y. & Skolnick J. (2004) Automated structure prediction of weakly homologous proteins on a genomic scale. Proc Natl Acad Sci U S A 101, 7594-7599.
8. Zhang Y., Kolinski A. & Skolnick J. (2003) TOUCHSTONE II: A new approach to ab initio protein structure prediction. Biophys J 85, 1145-1164.
9. Swendsen R.H. & Wang J.S. (1986) Replica Monte Carlo simulation of spin glasses. Physical Review Letters 57, 2607-2609.
10. Zhang Y., Kihara D. & Skolnick J. (2002) Local energy landscape flattening: Parallel hyperbolic Monte Carlo sampling of protein folding. Proteins 48, 192-201.
11. Jones D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292, 195-202.
12. Zhang Y., Hubner I., Arakaki A., Shakhnovich E. & Skolnick J. (2006) On the origin and completeness of highly likely single domain protein structures Proc Natl Acad Sci U S A 103, 2605-10.
13. Chen H. & Zhou H.X. (2005) Prediction of solvent accessibility and sites of deleterious mutations from protein sequence. Nucleic Acids Res 33, 3193-9.
14. Zhang Y. & Skolnick J. (2004) SPICKER: A clustering approach to identify near-native protein folds. J Comput Chem 25, 865-71.
15. Zhang Y. & Skolnick J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 33, 2302-2309.
16. Feig M., Rotkiewicz P., Kolinski A., Skolnick J. & Brooks C.L., 3rd. (2000) Accurate reconstruction of all-atom protein representations from side-chain-based low-resolution models. Proteins 41, 86-97.
17. Canutescu A.A., Shelenkov A.A. & Dunbrack R.L., Jr. (2003) A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci 12, 2001-14.

ZIB-Theseus - 409 models for 89 3D targets

An Automated Pipeline For Structure Prediction

P. May and T. Steinke

Zuse Institute Berlin, Germany

{patrick.may,steinke}@zib.de

The first step of our CASP7 structure prediction pipeline[1] was to identify domains within the query sequence. During CASP7 this was done using the Meta-DP server[2]. Then for all detected domain sequences suitable templates for homology modeling had to be found. A pipeline was established to perform successive PSI-Blast[3] searches automatically in order to find template structures. If no suitable template structure was found in the Protein Data Bank[4] (PDB), a PSI-Blast search in UniProt[5] was performed to initiate a new search in the PDB starting from the UniProt hits.

Starting with the templates found with the Blast searches, for every template 20 homology models were build with MODELLER[6] in parallel using additional Smith-Waterman alignments between query and template sequence, sequence

conservation information retrieved from the PSI-Blast profiles and secondary structure restraints from the DSSP[7] assignments for the template structures as input. The five best models according to the MODELLER score over all template structures were submitted.

If in the template search procedure described above no suitable template structure was found, a protein threading procedure using the THESEUS[8] implementation was initiated. THESEUS is a MPI-parallelized implementation of a protein threading based on a branch-and-bound search algorithm to find the optimal threading through a library of template structures. The template fold library was built on SCOP[9] version 1.69 domains. THESEUS uses a template core model based on secondary structure definition and a scoring function based on pseudo energies that include pairwise contacts, solvent accessibility, sequence profiles for query and template, variable gap lengths, and secondary structure matching between template and target as predicted by PsiPred[10]. From the highest scoring templates we selected the 20 most significant templates for further processing.

The reconstructed loops between the aligned adjacent template secondary structure elements were modeled with the inhouse developed aLip tool based on a comprehensive compilation of loop backbone conformations from a recent version of the PDB. The loop candidates were evaluated according to the RMSD between the stem atoms of template structure and loop candidate, sequence conservation, sequence properties and spatial constraints. Side chain modeling was done using MODELLER. Then a energy minimization was performed using MODELLER again generating 10 models for every template structure in parallel and submitting the best models according the MODELLER score.

1. May P., Ehrlich H.C., Steinke T. (2006) ZIB Structure Prediction Pipeline: Composing a Complex Biological Workflow through Web Services. In Euro-Par 2006 Parallel Processing, LNCS, 4128, 1148-1158.
2. Saini H.K., Fischer D. (2005) The Meta-DP: domain prediction meta-server. Bioinformatics 21, 2917-2920.
3. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Acids Res. 25, 3389-3402.
4. Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E. (2000) The Protein Data Bank. Nucleic Acids Res. 28, 235-242.
5. Bairoch A., Apweiler R., Wu C.H., Barker W.C., Boeckmann B., Ferro S., Gasteiger E., Huang H., Lopez R., Magrane M., Martin M.J., Natale D.A., O'Donovan C., Redaschi N., Yeh L.S. (2005) The Universal Protein Ressource (UniProt). Nucleic Acids Res. 33, D154-159.
6. Sali A., Blundell T.L (1993) Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234, 779-815.
7. Kabsch W., Sander C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2577-2637.
8. May P., Steinke T. (2006) THESEUS - Protein Structure Prediction at ZIB. ZIB Technical Report 06-24.
9. Lo Conte L., Brenner S.E., Hubbard T.J.P., Chothia C., Murzin A. (2002) SCOP database in 2002: refinements accommodate structural genomics. Nucl. Acids Res. 30, 264-267.
10. Jones D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292, 195-202.

Abstracts: Author Index

A

Affonnikov, 110
Agrawal, 106
Akiyama, 3
Ali, 51
Amemiya, 45
Antczak, 82
Arnautova, 107
Avbelj, 5

B

Bachinsky, 110
Baker, 6, 7, 8, 9, 97, 98, 119
Bakulina, 110
Baldi, 2, 11, 33, 47, 89
Baraniak, 4
Baratian, 79
Bates, 10
Baù, 33
Bazzoli, 38
Benkert, 108
Bennett-Lovsey, 112
Bhat, 7
Bhushan, 106
Bickerton, 114
Birzele, 70
Björklund, 105
Błażewicz, 82
Bliss, 117
Blundell, 59, 114
Bogatyreva, 92
Bohannan, 117
Boniecki, 50
Borreguero, 77, 113
Bose, 106
Bradley, 7
Brooks, 16
Brown, 114
Brylinski, 31
Bryson, 60, 61
Bu Do., 94
Bugalho, 48
Bujnicki, 4, 50
Burke, 59, 114

C

Camproux, 39
Cascone, 91
Cestaro, 19
Chaleil, 10
Chanco, 3
Chen C.Y., 13, 14
Chen H., 25, 77, 113
Chen Ji., 16
Chen M., 73, 74, 75
Cheng Ji., 2, 11, 33, 47
Cheng T., 114
Chi M-W, 64
Chinchio, 107
Chivian D., 6, 8, 9, 97, 98
Cho K-H, 112
Chodera, 32
Choi H-S, 81
Clark, 59
Colombo, 38
Cordes, 32
Coutsias, 32
Crivelli, 90
Cross, 115
Csaba, 70
Czaplewski, 107
Czerwoniec, 4
Czwojdrak, 4

D

Daisuke, 41, 49, 121
Daniluk, 67
Danzer, 118
Das, 6, 7, 97
Davuluri, 102
Debe, 118
Del Carpio, 17
Dematte, 19
Deronne, 62
Derreumaux, 39
Dill, 32
Ding F., 35, 36
Do S., 102
Dokholyan, 35, 36
Dolan, 115
Dong Q-W, 55
Duan Y., 37
Dudek, 57
Dunker, 58

E

Ekman, 105
Elber, 71
Elofsson, 105
Endou, 17

F

Fang Q., 109
Fang Yi., 20
Feder, 50
Fedorov, 65
Feig, 44
Fidelis, 67
Fienup, 118
Fijalkowski, 50
Finkelstein, 92
Fischer, 57, 77
Fitzjohn, 10
Floudas, 46
Fogolari, 85
Fontana, 17, 19
Friesner, 56
Fuchigami, 45
Furnham, 114

G

Gajda, 50
Galzitskaya, 83, 92
Gao Xi., 94, 96
Garbuzynskiy, 83, 92
Gattie, 57
Gerloff, 51
Goldblum, 114
Gopal, 87
Gordon, 114
Greenshpan, 63
Gront, 72
Gunda, 24
Gupta, 114
Guyon, 39

H

Han M., 80
Han W-S, 80
Hansmann, 78
Harris, 102
Harrison, 84
Hawkins, 64, 87
He Y., 83
Hegler, 86
Herbert, 112
Hirokawa, 3
Hirose, 22
Holm, 52
Honig, 56
Hori, 99
Horton, 3
Hosoi, 26, 27, 28, 29, 40, 41, 42, 43, 49, 69, 121
Hsieh, 3

Hsu C., 14, 102
Hung L-H, 103, 104
Hvidsten, 67

I

Ichiishi, 17
Imai, 66
Inoue, 21, 22
Ishida, 40, 65
Ivanisenko, 110
Ivankov, 92
Iwadate, 26, 27, 28, 29, 40, 41, 42, 43, 49, 69, 121

J

Jaś kowski, 82
Jayaram, 106
Jin W., 101
Jing Ju., 20
Jones, 60, 61
Joo K., 68

K

Kakuta, 12
Kalisman, 63
Kaminska, 4
Kaminski, 50
Kanabar, 102
Kanai, 22
Kanou, 26, 27, 28, 29, 40, 41, 42, 43, 49, 65, 69, 121
Karakaş, 76
Karmali, 114
Karplus, 102, 103
Karypis, 62
Kauffman, 62
Kaź mierkiewicz, 107
Keasar, 63
Kelley, 112
Khalili, 107
Khare, 6
Khatib, 102
Kihara, 25, 64, 87
Kim D., 6, 8, 9, 97, 98
Kim S-Y, 112
Kim T-K, 112
Kim Y-S, 64
Kinoshita, 40
Kitaura, 65
Klenin, 87
Kmiecik, 72
Kochanczyk, 31
Koike, 45
Kolinski, 72
Komorowski, 67
König, 51
Kosinski, 50
Koyama, 17
Kraut, 120
Krieger, 120
Kryshtafovych, 67
Kubo, 17
Kurcinski, 72
Kurowski, 50
Kuziemko, 56
Kwac, 86

L

La D., 64
Latek, 72
Lee Ji., 68
Lee J-K, 64, 81
Lee Jo., 68
Lee Ju., 112
Lee M-K, 80
Lee S., 68, 77
Lee S-J, 68
Lee S-Y, 113
Lei L., 55
Levitt, 69
Li Ji., 73, 74, 75
Li Mi., 94, 96
Li Sh., 94
Lise S., 60
Litvinov, 92
Liu P., 56
Liu S., 111
Liu T., 103, 104
Liwo, 107
Lobanov, 92
Lobley, 60
Lu M., 73
Luban, 64
Luethy, 72
Łukasiak, 82
Lund O., 30
Lundegaard, 30
Luo R., 3
Lupas, 11

M

Ma Ji., 73, 74, 75
Madera, 102
Maksimiak, 44
Mallick, 52
Malmstrom, 6, 7, 9, 98
Marko, 120
Martin, 33, 34, 39
Maupetit, 39
May, 124
McAllister, 46
McGuffin, 37, 60, 61, 81, 92
Meiler, 76
Meinke, 78
Miłostan, 82
Mitaku, 66
Miyamoto, 17
Mizuguchi, 114
Mohamed, 17
Mohan, 59
Mohanty, 78
Mokrab, 114
Montalvao, 114
Moon, 80
Mooney, 33
Morita, 12
Motono, 3
Munoz, 50
Murarka, 107

N

Nakamura, 12
Narang, 106
Nelson, 90
Nemoto, 114
Ngan S-C, 103, 104
Nie, 35, 36
Nielsen, 30
Noble, 119
Noguchi, 21, 22
Nunez, 114

O

Obarska, 50
Obradovic, 58
Offman, 10
Ohta, 26, 27, 28, 29, 40, 41, 42, 43, 49, 69, 121
Okazaki, 99
Ołdziej, 107
Oliveira, 48
Oomori, 45
Orlowski, 50
Ota, 3, 66
Ozkan, 32

P

Palik, 82
Palmer, 118
Pan, 83
Pandey, 106
Pandit, 77, 113
Papaj, 50
Park, 99, 101, 114
Pawlowski, 50
Peng, 58
Petrey, 56
Pettitt, 60
Pietal, 50
Pillardy, 71
Poleksic, 118
Pollastri, 19, 33, 34
Prentiss, 86
Priego, 114
Punta, 88

Q

Qian B., 6, 97
Qiu Ji., 71, 119
Qu Xi., 117

R

Radivojac, 58, 59
Raghava, 93
Rajgaria, 46
Randall, 2
Rangwala, 62
Rechtsteiner, 59
Ripoll, 23
Ritterson, 32
Rojas, 107
Rossi, 15
Rost, 88
Roterman, 31
Roytberg, 92

S

Saberi, 79
Sadeghian, 79
Sadowski, 60
Sahu, 106
Saini, 77
Samudrala, 103, 104
Sanborn, 102
Sasai, 66
Sasaki, 66
Sasin, 50
Sato, 3
Scheraga, 107
Schomburg, 108
Seo J., 68
Serohijos, 35, 36
Seth, 57
Shackelford, 102, 103
Sharikov, 114
Sharikova, 114
Sharma, 35, 36
Sheffler, 6, 119
Shell, 32
Shen H., 107
Shenoy, 106
Shi Ji., 118
Shimizu, 12, 22
Shirts, 56
Shortle, 109
Sirocco, 17
Skolnick, 77, 113
Smith, 114, 115
Soding, 11, 53, 54
Solovyev, 110
Song J-S, 64
Spasic, 107
Sraman, 6
Steinke, 124
Sternberg, 112
Stumpff-Kane, 44
Su C.T., 13, 14
Sumikoshi, 12
Summa, 69
Sundararajan, 24
Swanson, 116, 117
Sweredoski, 2, 33, 47

T

Takaba, 17
Takada, 99, 101
Takaya, 26, 28, 29, 40, 41, 42, 43, 49, 69
Takeda-Shitaka, 26, 27, 28, 29, 40, 41, 42, 43, 49, 65, 69, 121
Takizawa, 21
Tan Y.H., 25
Tanaka, 27, 121
Tang C.L., 56
Tanramluk, 114
Terashi, 26, 27, 28, 29, 40, 41, 42, 43, 49, 69, 121
Tettamanzi, 38
Thiltgen, 102
Thompson, 7
Thukral, 106
Titov, 110
Tkaczuk, 50
Tomii, 3, 21, 47, 48, 66
Toppo, 17, 19
Tosatto, 17, 18, 19, 33
Tsai J., 116, 117
Tsuboi, 17
Tsuji K., 66
Tuffery, 39
Tyka, 7

U

Umeyama, 26, 27, 28, 29, 40, 41, 42, 43, 49, 65, 69, 121
Urbič, 5

V

Vallat, 71
Velasco, 17, 19
Verma, 87
Vernon, 7
Vila, 107
Villani, 91
Voelz, 32
Vorobjev, 110
Vucetic, 58
Vullo, 19, 33, 34

W

Wallner, 105
Walsh, 33
Wang T., 37
Wang X-L, 55
Weinkam, 86
Wenzel, 87
West, 56
Wilton, 52
Wollacott, 6
Wolynes, 86
Wong C., 102
Worth, 114
Wu G.A., 32
Wu Yi., 73, 74, 75
Wymore, 120

X

Xie L., 56
Xu Ji., 94, 96

Y

Yang Y.F., 25
Yin S., 35, 36
Yoon C-N, 64, 80, 81
Yu Li., 94
Yu M., 92

Z

Zauli, 15
Zhang C., 111
Zhang Y., 122, 123
Zhao S., 56
Zhou H., 77, 113
Zhou Y., 111
Zhu J., 56
Zhu K., 56
Zimmer, 70
Zimmermann, 78
Zong C., 86

Abstracts: Group Index

3DPRO - 500 MODELS FOR 100 3D TARGETS 2

FOLDPRO - 600 MODELS FOR 100 3D/100 DP TARGETS 2

ABIPRO (SERVER, 3D) - 495 MODELS FOR 99 3D TARGETS 2

3D Structure Prediction Using FOLDpro, 3Dpro, and ABIpro 2

AMBER/PB - 96 MODELS FOR 1 3D/ 91QA TARGETS..... 3

Quality Assessments of Server Results with AMBER/PBSA 3

ANDANTE - 552 MODELS FOR 100 3D/ 42 DP/3 FN/1QA TARGETS..... 3

Tertiary Structure Prediction of CASP7 Targets Using Exhaustive Modeling and Evaluation 3

AMU-BIOLOGY - 322 MODELS FOR 92 3D/ 19 FN TARGETS..... 4

Combination of template-based and template-free modeling 4

AVBELJ - 22 MODELS FOR 7 3D TARGETS 5

Predictions of three-dimensional structures of proteins using Monte Carlo simulations and electrostatic screening model 5

BAKER - 533 MODELS FOR 99 3D/8 TR TARGETS..... 6

Template-based Structure Prediction in CASP7 by Rosetta and Rosetta@home 6

BAKER - 533 MODELS FOR 99 3D/8 TR TARGETS..... 7

Protein structure prediction by free modeling and Rosetta@home in CASP7..... 7

BAKER-DP_HYBRID - 100 MODELS FOR 100 DP TARGETS 8

Hybrid domain parsing with Ginzu and RosettaDOM..... 8

BAKER-ROSETTADOM - 99 MODELS FOR 99 DP TARGETS 9

The RosettaDOM Domain Parsing Protocol 9

BATES - 536 MODELS FOR 100 3D/ 9TR TARGETS..... 10

3D-JIGSAW - 536 MODELS FOR 100 3D/ 9TR TARGETS..... 10

3D-JIGSAW-RECOMB - 462 MODELS FOR 100 3D TARGETS 10

3D-JIGSAW-POPULUS - 500 MODELS FOR 100 3D TARGETS 10

Using genetic algorithms to recombine and refine protein models 10

BAYESHH - 100 MODELS FOR 100 3D TARGETS..... 11

Homology-based structure prediction by HMM-HMM comparison and stochastic alignment sampling..... 11

BETAPRO - 100 MODELS FOR 100 RR TARGETS 11

SVMCON - 100 MODELS FOR 100 RR TARGETS 11

(SERVER, CONTACT) 11

Contact Map Prediction Using BETApro and SVMcon..... 11

BILAB - 619 MODELS FOR 100 3D/100 QA/ 8 TR TARGETS 12

BILAB-ENABLE - 434 MODELS FOR 99 3D TARGETS 12

Automated tertiary structure prediction of proteins using fold recognition, model quality assessment, and fragment assembly... 12

BIME@NTU - 196 MODELS FOR 98 DR/ 98 RR TARGETS 13

DisorderPSC: Protein Disorder Prediction by Condensed PSSM, Secondary Structure, and Conservation Information..... 13

BIME@NTU - 196 MODELS FOR 98 DR/ 98 RR TARGETS 14

Prediction of Remote Residue Contacts by Concurrent Sequence Conservation..... 14

BIODEC - 66 MODELS FOR 65 3D TARGETS 15

All-atom Models Starting from Entropy-filtered Alignments 15

BROOKS_CASPR - 108 MODELS FOR 23 3D/ 7 TR TARGETS..... 16

High-Resolution Structure Refinement Using Implicit Solvent and Replica Exchange 16

CADCMLAB - 476 MODELS FOR 96 3D TARGETS..... 17

Combining Spectral Based Sequence Comparison Methods with Orthodox Sequence Alignment Techniques for Protein Fold Recognition and 3-D Structure Prediction..... 17

CASPITA-FOX - 499 MODELS FOR 100 3D TARGETS 17

FOX (FOld eXtractor): A protein fold recognition method using iterative PSI-BLAST searches and structural alignments 17

CASPITA-FRST - 93 MODELS FOR 93 QA TARGETS..... 18

The Victor/FRST Function for Model Quality Estimation..... 18

CASPITA-FRST-SVM - 98 MODELS FOR 98 QA TARGETS.....19

FRST-SVM: Predicting Model Quality from Statistical Potentials and Structural Features Using Kernel Machines..... 19

CASPITA-GORET - 227 MODELS FOR 100 FN TARGETS.....19

GOretriever (Gene Ontology retriever): a fast automated protein function annotation based on semantic similarities 19

CBIS - 15 MODELS FOR 4 3D TARGETS 20

A new ab initio mathematical model for protein structure predictions..... 20

CBRC-DP_DR - 200 MODELS FOR 100 DP/ 100 DR TARGETS 21

Prediction of disordered coil regions in proteins by fold recognition and secondary structure prediction 21

CBRC-DR - 100 MODELS FOR 100 DR TARGETS.....22

POODLE: predicting protein disorder using machine-learning approaches..... 22

CBSU - 287 MODELS FOR 100 3D TARGETS 23

Protein Structure Models based on Fold-Recognition Templates and Their Remote Structural Neighbors 23

CDAC - 4 MODELS FOR 4 3D TARGETS.....24

Hybrid Methods for Predicting the Protein Structures..... 24

CHEN-TAN-KIHARA - 653 MODELS FOR 97 3D/ 34 QA TARGETS 25

Fold recognition prediction based on suboptimal DP..... 25

CHEN-TAN-KIHARA-QA - 653 MODELS FOR 97 3D/ 34 QA TARGETS 25

Quality assessment using the diversity of suboptimal alignments25

CHIMERA - 542 MODELS FOR 100 3D TARGETS.....26

Protein Structure Prediction using SKE-CHIMERA..... 26

CIRCLE - 500 MODELS FOR 100 3D TARGETS.....27

CIRCLE: Full automated homology-modeling server using the 3D1D scoring functions 27

CIRCLE-FAMS - 496 MODELS FOR 100 3D TARGETS.....28

Selection from all the server models using original 3D 1D program -"CIRCLE" 28

CIRCLE-QA - 100 MODELS FOR 100 QA TARGETS.....29

CIRCLE for quality assessment in CASP7.....29

CPHMODELS - 49 MODELS FOR 49 3D TARGETS.....30

CPHmodels 30

CRACOW.PL - 58 MODELS FOR 52 3D TARGETS.....31

Simulation of protein folding process rather then protein structure prediction.....31

DILL-ZAP - 30 MODELS FOR 6 DR TARGETS 32

Physics-Based Protein Folding by Zipping and Assembly.....32

DISPRO (SERVER, DISORDER) - 100 MODELS FOR 100 DR TARGETS 33

Protein Disordered Region Prediction Using DISpro.....33

DISTILL - 800 MODELS FOR 100 3D/100 DP/100 DR/100RR TARGETS33

DISTILL_HUMAN - 800 MODELS FOR 100 3D/100 DP/100 DR/100RR TARGETS 33

Draft protein structures by machine learning.....33

DISTILLFM - 68 MODELS FOR 68 RR TARGETS 34

A Filtering Approach for Improved Modeling of Predicted Contact Maps 34

DOKHLAB - 114 MODELS FOR 26 3D TARGETS 35

Ab Initio Structure Prediction by Rapid Sampling of Protein Conformational Space 35

DOKHLAB - 114 MODELS FOR 26 3D TARGETS 36

Atomic-Resolution Prediction of Protein Structure Using Constrained Replica-Exchange Annealing.....36

DOMFOLD- 100 MODELS FOR 100 DP TARGETS 37

A combined approach to automated protein domain prediction...37

DUAN_GROUP- 13 MODELS FOR 7 TR TARGETS 37

Protein Structure Refinement by Molecular Dynamics Simulation 37

EATORP- 20 MODELS FOR 20 3D TARGETS 38

Ab Initio Protein Structure Prediction with a Dipeptide-Assembly Evolutionary Algorithm 38

EBGM-LBT- 40 MODELS FOR 15 3D TARGETS 39

Assessing a new approach for protein structure modeling combining structural alphabet local conformation prediction and greedy algorithm for reconstruction 39

FAIS - 338 MODELS FOR 78 3D/100 DR TARGETS 40

Protein tertiary structure prediction based on contact number prediction 40

FAMS - 500 MODELS FOR 100 3D TARGETS 40

Automatic modeling server using homology modeling method and *ab-initio* method 40

FAMS_ACE - 500 MODELS FOR 100 3D TARGETS 41

Model selection from server results using original threading(3D1D) program and consensus 41

FAMSD - 500 MODELS FOR 100 3D TARGETS 42

Homology modeling server providing 42
side chain models with high accuracy 42

FAMS-MULTI- 509 MODELS FOR 100 3D/9 TR TARGETS 43

Homology modeling meta-server using multiple reference proteins 43

FEIG - 469 MODELS FOR 99 3D TARGETS 44

Sampling and Scoring Strategies in an Iterated Protocol for Protein Structure Prediction 44

FLEIL - 311 MODELS FOR 63 3D TARGETS 45

Comparative modeling with all-atom refinement using molecular dynamics simulation 45

FLOUDAS - 150 MODELS FOR 30 3D/1 TR TARGETS 46

First Principles Protein Structure Prediction 46

FOLDPRO (SERVER, DOMAIN) - 600 MODELS FOR 100 3D/ 47

100 DP TARGETS 47

Domain Prediction Using FOLDpro and DOMpro 47

FORTE1 - 499 MODELS FOR 100 3D TARGETS 47

FORTE1: A Profile-Profile Comparison Method for Fold Recognition 47

FORTE2 - 499 MODELS FOR 100 3D TARGETS 48

FORTE2: Automated Fold Recognition Server with Enhanced Profile Library 48

FPSOLVER_SERVER- 436 MODELS FOR 96 3D TARGETS 48

Ab-initio Protein Structure Prediction Using Backtrack Search 48

FUNCTION - 500 MODELS FOR 100 3D TARGETS 49

Building many models using FAMS and selecting model with Special scoring Function 49

GENESILICO - 616 MODELS FOR 96 3D/ 92 DP / 91 QA TARGETS 50

Identification and refinement of potential errors in protein structures 50

GERLOFF 51

A simplified representation of electrostatic model surfaces for addressing protein-protein interaction problems 51

GTG - 201 MODELS FOR 45 3D TARGETS 52

Transitive alignment of distantly-related proteins 52

HHPRED1 – 298 MODELS FOR 100 3D/98 DP TARGETS 53

Homology-based structure, function, and domain prediction by HMM-HMM comparison 53

HHPRED2 - 100 MODELS FOR 100 3D TARGETS 53

Homology-based structure prediction by HMM-HMM comparison and multiple template selection 53

HHPRED3 - 300 MODELS FOR 100 3D/100 DP/100 FN TARGETS.....54

Homology-based structure, function, and domain prediction by HMM-HMM comparison, multiple template selection, and intermediate profile search..... 54

HIT-ITNLP - 459 MODELS FOR 95 3D TARGETS.....55

The FRPPSP fold recognition method..... 55

HONIGLAB - 174 MODELS FOR 90 3D/2 TR TARGETS.....56

Template-based protein structure prediction using an automated modeling pipeline, manual target analysis and fast model evaluation..... 56

HPREDGRP- 100 MODELS FOR 100 3D TARGETS.....57

VERIFY- 100 MODELS FOR 100 3D TARGETS.....57

Fischerlab automatic predictions..... 57

IGOR - 66 MODELS FOR 52 3D TARGETS.....57

A Simple Easily-Integratable Model of a Protein Chain..... 57

ISTZORAN - 99 MODELS FOR 99 DR TARGETS.....58

Length-Dependent Prediction of Protein Intrinsic Disorder..... 58

IUB-INFO - 197 MODELS FOR 81 FN TARGETS.....59

Protein function prediction from sequence, properties and literature..... 59

JIVE - 458 MODELS FOR 94 3D/7 TR TARGETS.....59

Prediction of protein structure using template-free assembly of secondary and super-secondary motifs..... 59

JONES-UCL - 322 MODELS FOR 99 3D/83 QA/2 TR TARGETS.....60

Use of Fragment Assembly, Threading and Model Quality Assessment Methods to Predict Protein Folds..... 60

JONES-UCL (SERVERS) - 322 MODELS FOR 99 3D/83 QA/2 TR TARGETS.....61

Protein Structure Prediction Servers at UCL..... 61

KARYPIS - 161 MODELS FOR 83 3D TARGETS.....62

Protein Structure Prediction using learning based methods, fragment assembly and simple alignment techniques.....62

KEASAR - 573 MODELS FOR 100 3D/84 QA/2 TR TARGETS.....63

Refinement of Fold Recognition Models with Cooperative Solvation Potentials.....63

KIHARA_PFP - 99 MODELS FOR 97 FN TARGETS.....64

Partially automated, comprehensive annotation with PFP.....64

KIST - 450 MODELS FOR 97 3D TARGETS.....64

Simulation of Protein Folding Structures.....64

KITAURA-FAMS - 8 MODELS FOR 8 3D TARGETS.....65

Refinement of protein structures using the fragment molecular orbital (FMO) method.....65

KORO - 200 MODELS FOR 40 3D TARGETS.....66

A coarse-grained Langevin molecular dynamics approach to de novo structure prediction.....66

LARGO - 28 MODELS FOR 1 3D/27 QA TARGETS.....66

Quality Assessment of 3D-models by LIBRA_rotamer.....66

LCBDAVIS - 91 MODELS FOR 91 FN TARGETS.....67

Prediction of protein function using local descriptors of protein structure.....67

LEE - 1098 MODELS FOR 99 3D/99 DP/99 QA/9 TR TARGETS.....68

A Template based Modeling based on Global Optimization.....68

LEVITT - 8 MODELS FOR 8 TR TARGETS.....69

Pairwise Atomic Potentials and Near-Native Structure Refinement: *In vacuo* energy minimization.....69

LIGAND-CIRCLE - 439 MODELS FOR 88 3D TARGETS.....69

Ligand-Circle: CIRCLE to evaluate binding sites.....69

LMU - 191 MODELS FOR 65 3D/98 DP TARGETS.....70

Fold recognition and Alignment optimization using.....70
AutoSCOPe and protein class flexibility.....70

LOOPP - 500 MODELS FOR 100 3D TARGETS....71

Learning, Observing and Outputting Protein Patterns (LOOPP).. 71

LTB-WARSAW - 397 MODELS FOR 83 3D/1 TR TARGETS....72

Automated approach to protein structure prediction with the lattice reduced model and BioShell toolkit suite 72

LUETHY - 100 MODELS FOR 100 DP TARGETS....72

Consensus distance matrices derived from server predictions used as folding potential.... 72

MA-OPUS-QUALITY ASSESSMENT - 702 MODELS FOR 99 3D/100 DP/99 QA TARGETS....73

Model Quality Assessment Based on a Novel Cα-based Empirical Potential 73

MA-OPUS-DE NOVO - 702 MODELS FOR 99 3D/100 DP/99 QA TARGETS....74

OPUS: A New *De Novo* Protocol for Determining Overall Protein Topology 74

MA-OPUS-DOM - 106 MODELS FOR 99 DP TARGETS....75

OPUS-DOM: A Novel Method for Domain Boundary Prediction in CASP7.... 75

MEILER - 97 MODELS FOR 97 PR TARGETS....76

Contact Prediction Using Artificial Neural Networks.... 76

META-DP - 100 MODELS FOR 100 DP TARGETS....77

Meta-DP: Domain Prediction Meta Server.... 77

METATASSER - 929 MODELS FOR 100 3D TARGETS....77

MetaTasser: A 3D-jury threading approach with TASSER model assembly/refinement 77

MTUNIC - 483 MODELS FOR 97 3D TARGETS....78

Parallel Tempering Monte Carlo based Protein Structure prediction and Refinement 78

MUMSSP - 19 MODELS FOR 13 3D TARGETS....79

Loop refinement and geometry optimization: key steps in protein modeling 79

NANO3D - 316 MODELS FOR 64 3D TARGETS....80

Ab initio protein folding.... 80

NANODESIGN - 322 MODELS FOR 82 3D/1 TR TARGETS....80

Sidechain optimization using NanoDesign.... 80

NANOMODEL - 492 MODELS FOR 100 3D TARGETS....81

NanoModel: Protein structure modeling pipeline.... 81

NFOLD - 500 MODELS FOR 100 3D TARGETS....81

Fully automated protein fold recognition using a modified version of the nFOLD protocol.... 81

NN_PUT_LAB - 279 MODELS FOR 94 3D/ 93 DP TARGETS....82

DomAnS method – the new approach used for predicting domains boundaries in proteins.... 82

NN_PUT_LAB - 279 MODELS FOR 94 3D/ 93 DP TARGETS....82

3D Judge – meta predictor for 3D protein structure.... 82

OKA - 206 MODELS FOR 4 3D/100 DP/99 DR TARGETS....83

Entropy capacity determines protein folding rate.... 83

PAN - 586 MODELS FOR 100 3D/73 FN TARGETS....83

'Threading' with structural profile 83

PANTHER - 222 MODELS FOR 66 3D TARGETS....84

Alignment and Regularization in Modeling 84

PC2CA - 100 MODELS FOR 100 QA TARGETS....85

PC2CA: a pseudocovalent model for protein structures.... 85
with two centers of interaction per amino acid.... 85

PETER-G-WOLYNES - 160 MODELS FOR 32 3D TARGETS....86

Associative Memory Hamiltonian Protocol for CASP7.... 86

PFP_HAWKINS - 36 MODELS FOR 36 FN TARGETS....87

Fully automated GO term prediction with PFP.... 87

POEM-REFINE - 135 MODELS FOR 27 3D TARGETS.....87

De novo protein structure prediction by all-atom..... 87
free-energy refinement with PFF01..... 87

PROFCON-ROST - 77 MODELS FOR 77RR TARGETS.....88

Prediction of protein residue internal contact through Neural Networks..... 88

PROTEINSHOP - 41MODELS FOR 9 3D TARGETS.....90

Protein Structure Prediction Using ProteinShop..... 90

PROTEO - 63 MODELS FOR 62 3D TARGETS.....91

Protein Folding Simulations through low and high resolution models..... 91

PUSHCHINO - 4 MODELS FOR 4 3D TARGETS.....92

Combining sequence alignment tools with threading approach to improve the quality of protein structure prediction..... 92

QA-MODFOLD - 100 MODELS FOR 100 QA TARGETS.....92

ModFOLD: a consensus of model quality assessment programs using an artificial neural network..... 92

RAGHAVA-GPS-MANGO - 285 MODELS FOR 95 FN TARGETS.....93

MANGO: prediction of Genome Ontology (GO) class of a protein from its amino acid and dipeptide composition using nearest neighbor approach..... 93

RAPTOR-ACE - 500 MODELS FOR 100 3D TARGETS.....94

An integer linear programming based..... 94
consensus fold recognition method..... 94

RAPTORESS - 500 MODELS FOR 100 3D TARGETS.....96

An Atom-level Refinement Approach for..... 96
Protein Structure Prediction..... 96

ROBETTA - 495 MODELS FOR 99 3D TARGETS.....97

Robetta *De Novo* and Homology Modeling in CASP7..... 97

ROBETTA-GINZU - 211 MODELS FOR 99 DP TARGETS.....98

Ginzu homolog identification and domain parsing in CASP7..... 98

ROKKO - 476 MODELS FOR 98 3D TARGETS.....99

Template-free Prediction by Fragment Assembly with SimFold Energy Function at CASP7.....99

ROKKY - 444 MODELS FOR 98 3D TARGETS.....101

De novo Structure Prediction Server by Fragment Assembly with SimFold Energy Function.....101

SAM-T06 - 682 MODELS FOR 100 3D/93 RR/ 7 TR TARGETS.....102

SAM-T06: Full 3D predictions from UCSC.....102

SAM-T06 - 682 MODELS FOR 100 3D/93 RR/ 7 TR TARGETS.....103

Residue-Residue Contact Prediction Using Selected Correlation Statistics.....103

PROTINFO - 500 MODELS FOR 100 3D TARGETS.....103

SAMUDRALA - 611 MODELS FOR 99 3D /5 FN /99 QA/.....103

4 TR TARGETS.....103

Comparative model refinement using graph-theoretic and consensus-based restrained molecular dynamics approaches (PROTINFO/SAMUDRALA).....103

SAMUDRALA-AB - 493 MODELS FOR 99 3D TARGETS.....104

Constraint-based free modeling.....104

SBC - 500 MODELS FOR 100 3D TARGETS.....105

Automatic predictions of protein structure, quality assessments, local residue based quality and function from Stockholm University. 105

SCFBIO-IITD - 20 MODELS FOR 3 3D TARGETS.....106

Bhageerath: An Energy Based Protein Tertiary Structure Prediction Server for Small Globular Proteins.....106

SCHERAGA - 220 MODELS FOR 43 3D TARGETS.....107

Physics-based protein-structure prediction using mesoscopic dynamics and the Conformational Space Annealing (CSA) method with the UNRES force field - test on CASP7 targets.....107

SCHOMBURG-GROUP - 133 MODELS FOR 18 3D/65 QA TARGETS. 108

A comparative modeling pipeline combined with a statistical potential scoring function..... 108

SHORTLE - 401 MODELS FOR 91 3D/5 TR TARGETS..... 109

Homology modeling with Atom-Based Statistical Potentials and a Simple Genetic Algorithm 109

SOFTBERRY - 196 MODELS FOR 96 3D/100 DR TARGETS..... 110

Softberry tools for protein structure analysis and modeling..... 110

SP4 - 500 MODELS FOR 100 3D TARGETS..... 111

Template-based Protein Structure Prediction by SP^4..... 111

SSU - 125 MODELS FOR 25 3D TARGETS..... 112

Protein tertiary structure prediction using ECEPP/SM potential energy function and Monte Carlo with minimization..... 112

STERNBERG - 190 MODELS FOR 99 3D TARGETS..... 112

Integrating *ab* initio folding, domain boundary prediction and in-house ensemble fold recognition in Phyre..... 112

TASSER - 1027 MODELS FOR 100 3D/100 QA/8 TR TARGETS..... 113

TASSER for protein structure prediction in CASP7 113

TENETA - 225 MODELS FOR 98 3D TARGETS..... 114

TENETA - HMM-oriented Structure Prediction Method..... 114

TLBGROUP - 20 MODELS FOR 14 3D/ 1 FN TARGETS..... 114

Combining homology recognition and knowledge based modelling with ensembl generation. 114

TRIPOS-CAMBRIDGE - 13 MODELS FOR 19 3D TARGETS..... 115

ORCHESTRAR Homology Modeling..... 115

TSAILAB - 245 MODELS FOR 42 3D/7 TR TARGETS..... 116

Measuring 3D information from protein structure: "3D-bits", an intuitive yet quantitative assessment of comparative modeling predictions..... 116

TSAILAB - 245 MODELS FOR 42 3D/7 TR TARGETS..... 117

Side-Chain Guided Protein Refinement 117

UCB-SHI - 885 MODELS FOR 96 3D/98 QA TARGETS..... 118

Modeling and Quality Assessment using HARMONY3 and QUAD 118

UNI-EID_BNMX - 462 MODELS FOR 100 3D TARGETS..... 118

UNI-EID_EXPM - 100 MODELS FOR 100 3D TARGETS..... 118

UNI-EID_SFST – 454 MODELS FOR 100 3D TARGETS..... 118

A probabilistic approach to remote homology detection and 3D protein structure modeling..... 118

UWSCORE - 97 MODELS FOR 97 QA TARGETS..... 119

Protein Structure Quality Prediction with Support Vector Regression 119

WYMORE- 104 MODELS FOR 37 3D TARGETS..... 120

Comparative modeling using a stochastic alignment algorithm and statistical potentials..... 120

YASARA - 68 MODELS FOR 21 3D/21 QA/5 TR TARGETS..... 120

High resolution refinement with YASARA 120

YU-BA - 72 MODELS FOR 72 FN TARGETS..... 121

Yuba (Yet Unnamed Binding site Assessment) aiming at the binding site prediction in protein function category of CASP7..... 121

ZHANG - 500 MODELS FOR 100 3D TARGETS..... 122

Protein structure prediction by iterative TASSER simulations... 122

ZHANG-SERVER - 500 MODELS FOR 100 3D TARGETS..... 123

Server prediction by iterative TASSER simulations in CASP7.. 123

ZIB-THESEUS - 409 MODELS FOR 89 3D TARGETS..... 124

An Automated Pipeline For Structure Prediction..... 124