CASP9 ABSTRACT BOOK

Critical Assessment of Techniques for Protein Structure Prediction

Ninth Meeting

PACIFIC GROVE, CALIFORNIA, USA

DECEMBER 5-9, 2010

TABLE OF CONTENT

3D-JIGSAW-4.0
3D-JIGSAW-4.5

3D-JIGSAW-4.0 & 3D-JIGSAW-4.5

R.A.G. Chaleil[1], M.N. Offman[1] and P.A. Bates[1]
1 Biomolecular Modelling Laboratory – Cancer Research UK London Research Institute
raphael.chaleil@cancer.org.uk

3D-JIGSAW-4.0 and 3D-JIGSAW-4.5 are modified versions of a previously reported genetic algorithm for template mixing[1]. There are notable differences between the two that we wished to investigate: differences in the initial template identification and selection; different protocols for crossover selection.

Methods

For both servers, to implicitly include sequence context dependent information, template identification is done using HHsearch[2], however, the Hidden Markov Database and the Hidden Markov query are generated using profiles from CSI-BLAST[3]. In addition, for version 4.5, extra alignments for selected templates are added from our in-house alignment protocol that mixes secondary structure and homologous sequence information.

A second variation between versions 4.0 and 4.5 resides in the selection of recombination hotspots in the Genetic Algorithm. In the first case, the recombinations are selected at random. In the second case, the two parent structures are first superimposed using a local implementation of the algorithm described by Gerstein and Levitt[5], then recombination points are selected at position in the superimposition where equivalent alpha carbon are less than two Angstroms apart.

Manual models were generated by the following protocol: the best five server models, according to ranking using the DFIRE[4] pair potential, were individually split into domains using an implementation of the Protein Peeling algorithm[6]. For each of the five models, fragment boundaries, mostly well-defined hinge regions, were adjusted by sampling the dihedral angles (Phi and Psi) of the five residues either sides of each hinge-point. This optimization was performed within the framework of a Particle Swarm Optimisation[7] algorithm, with the energy function and parameters were taken from DFIRE[4].

Results

From our own analysis of currently available Target solutions, 3D-JIGSAW-4.5 appears to perform better on the more difficult targets than version 4.0, and the converse for easier Targets. We attribute this mainly to version 4.5 sampling a more diverse fragment space.

Our manual intervention protocol is an early developmental algorithm to try to adjust top server models towards the Target structure by sampling key hinge-point regions within and between folds. As described above, the new method is based upon a PSO. The main difficultly we experienced was not being able to indentify the better server models from which to start the optimization. However, if we did select a high quality, or an ensemble of better server models, the new protocol showed some promise. For a few Targets, particularly those with multiple domains - hence often well-defined hinge points - we did manage to adjust backbone angles in the right direction.

Problems associated with both our automatic server and intervention protocols: the use of energy functions that are still not refined enough for the wide variety of conformations investigated; coupled to

this, is the need to sample conformational space even more deeply – crossover points and mutation frequencies. To perform such thorough sampling requires considerable computer resources, therefore, our investigations into efficient search algorithms such as the PSO may alleviate some of the pressure our genetic algorithm imposes on computer resources.

Availability

The server 3D-JIGSAW-Populus is available online at http://bmm.cancerresearchuk.org/~populus

1. Offman MN, Fitzjohn PW, Bates PA. Developing a move-set for protein model refinement. Bioinformatics. 2006; 22(15):1838-1845.Soding J. Protein homology detection by HMM-HMM. Bioinformatics. 2005; 21:951-960.Biegert A, Soeding J, Sequence context-specific profiles for homology searching. PNAS,2009 Mar 10;106(10):3770-5 Yang Y, Zhou Y,``Specific interactions for *ab initio* folding of protein terminal regions with secondary structures.", *Proteins* **72**, 793-803 (2008).
2. Gerstein M, Levitt M, Comprehensive assessment of automatic structural alignment against a manual standard, the scop classification of proteins. Protein Sci. 1998 Feb;7(2):445-56
3. Gelly JC, de Brevern AG, Hazout S, 'Protein Peeling': an approach for splitting a 3D protein structure into compact fragments. Bioinformatics. 2006 Jan 15;22(2):129-33. Kennedy, J.; Eberhart, R. (1995). "Particle Swarm Optimization". *Proceedings of IEEE International Conference on Neural Networks*. **IV**. pp. 1942–1948

3DLigandSite1-4

Using 3DLigandSite to making binding site predictions in CASP9

M.N. Wass, L.A. Kelley and M.J.E. Sternberg
Structural Bioinformatics group, Centre for Bioinformatics, Imperial College London
Mark.wass04@imperial.ac.uk, l.a.kelley@imperial.ac.uk, m.sternberg@imperial.ac.uk

3DLigandSite[1] (http://www.sbg.bio.ic.ac.uk/3dligandsite) is an automated method for the prediction of ligand binding sites. It was the developed as a result of our successful manual approach for prediction in CASP8[2]. In CASP9 we have incorporated new features into 3DLigandSite for the both server and human predictions. We ran multiple automated servers, which used different structural models as templates and different cut offs within the predictive process.

Methods

Full details of the 3DLigandSite algorithm are available in Wass et al., (2010). In brief 3DLigandSite uses Phyre[3] to model the structure of the target protein. The model is used to perform a structural search of a database of ligand-bound protein structures. This identifies similar structures to the target protein, which are aligned to the target using MAMMOTH[4]. This superimposes the ligands from the similar structures on to the target model. The ligands are clustered spatially and the largest cluster is focused on as the most likely binding site. The distance of residues from the ligands in the cluster is used to determine which residues are predicted to form part of the binding site in the target structure.

For CASP9 we incorporated the Jensen-Shannon divergence[5] conservation score into the 3DLigandSite approach. The conservation score was used to filter the predictions made from the clustered ligands described above, whereby any residues with a Jensen Shannon divergence score below a threshold were excluded from the prediction.

Servers 3DLigandSite1 and 3DLigandSite2 both used the standard Phyre server (http://www.sbg.bio.ic.ac.uk/phyre) to model the target structure. They also used different thresholds for for clustering and conservation. Servers 3DLigandSite3 and 3DLigandSite4 used the structure predictions of our Phyre2 CASP9 server and they also used different thresholds.

For human predictions a consensus approach was used. The CASP9 server predictions were downloaded and clustered using 3DJury[6]. The top 6 models (obtained from different groups) were individually run through 3DLigandSite. The results were manually combined. Individual residues were predicted to form part of the binding site based on the number of 3DLigandSite runs that had predicted them, their conservation score and on visualization of the modeled protein and the clustered ligands. Additional functional information for the targets was sought from UniProt[7], Pfam[8],Interpro[9] and ConFunc[10] to aid the manual process particularly to determine the likely ligands of the target.

Availability

3DLig and Site is available at http://www.sbg.bio.ic.ac.uk/3dligandsite

1. Wass,M.N., Kelley,L.A., & Sternberg,M.J. (2010) 3DLigandSite: predicting ligand-binding sites using similar structures. *Nucl. Acids Res.* **38**, W469-473.
2. Wass,M.N. & Sternberg,M.J. (2009) Prediction of ligand binding sites using homologous structures and conservation at CASP8. *Proteins* **77**, 147-151.
3. Kelley,L.A. & Sternberg,M.J. (2009) Protein structure prediction on the Web: a case study using the Phyre server. *Nat Protoc* **4**, 363-371.
4. Ortiz,A.R., Strauss,C.E. & Olmea,O. (2002) (MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci*, **11,** 2602-2621.
5. Capra,J.A. & Singh,M. (2008) Characterization and prediction of residues determining protein functional. *Bioinformatics* **24**, 1473-1480.
6. Ginalski,K., Elofsson,A., Fischer,D., & Rychlewski,L. (2003) 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics* **19**, 1015-1018.
7. The UniProt Consortium (2009) The Universal Protein Resource (UniProt) 2009. *Nucl. Acids Res.* **37**, D169-174.
8. Finn,R.D., *et al.* (2010) The Pfam protein families database. *Nucleic Acids Research* **38**, D211-D222.
9. Hunter,S., *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Research* **37**,D211-D215.
10. Wass,M.N. & Sternberg,M.J. (2008) ConFunc--functional annotation in the twilight zone. *Bioinformatics* **24**,798-806.

3SP-TsaiLab

A side-chain centric method for template-based structure prediction

R. Day[1], H. Joo[1], A. Chavan[1], K.P. Lennox[2], A. Chen[3], D.B. Dahl[2], M. Vannucci[3], and J.W. Tsai[1]
University of the Pacific, [2] - Texas A&M, [3] - Rice University
jtsai@pacific.edu

Current template based protein structure prediction methods take a template backbone, perturb it, and then rebuild the side chains. These methods have proven successful and have been extensively optimized over the last decade, to the point where formidable efforts are now required to realize modest improvements. We propose a new paradigm for template based modeling: place side chains, then model the backbone. Our prediction method consists of three new and unique components: a novel side-chain centric method of perturbing protein conformations (3SP), a Dirichlet process mixture of hidden Markov models of backbone φ-ψ space for loop modeling (Cortorgles), and a unique volume and torsion angle based scoring function (volangle score).

Methods

In 3SP, the maximal cliques from the contact graph for a template structure are aligned to maximal cliques[1] from all protein structures in the PDB. The contact graph is determined from the Delaunay tessellation of all protein heavy atoms[2]. Geometrically similar cliques are selected and used to build density estimations of the positions of the side-chain centers of mass. Draws on these distributions are used to perturb the positions of the side-chains. Conditional distributions of Cα positions are then created based on the new side-chain positions and draws from these distributions are used to perturb the protein backbone. In this way, the position of the side-chain dictates the position of the backbone. Regions of the protein that are not perturbed by 3SP (i.e. residues that do not participate in many contact cliques) are considered to be loop regions and are modeled using Cortorgles. Cortorgles uses a hidden Markov model that captures known properties of protein secondary structure and phi/psi information from the template(s) as a centering distribution for a Dirichlet process[3]. This allows us to model loop φ-ψ from sparse data. φ-ψ samples are converted to Cartesian coordinates using standard values for bond lengths and angles. All-atom models of the protein are built up from our Cα + center-of-mass models using Pulchra[4], and then scored. Our scoring function considers residue volumes calculated using Voronai polyhedra[5] and the φ, ψ, and χ1 angles of each residue. It is parameterized based on a large set of native state molecular dynamics simulations[6] and on decoy sets from previous CASP experiments.

Availability

The 3SP software and dataset are available by contacting the authors, as is the volangle score. The Cortorgles software is available online at: http://www.stat.tamu.edu/~dahl/software/.

1. Bron,C. & Kerbosch,J. (1973). Finding all cliques of an undirected graph. *Communications of the ACM*. **16**, 575-577.
2. Delaunay,B. (1934). Sur la sphere vide. *Izv Akad Nauk SSSR, Otdelenie Matematicheskikh i Estestvennykh Nauk.*. **7**, 793-800.
3. Lennox,K.P., Dahl,D.B., Vannucci,M., & Tsai,J. (2009). Density estimation for protein conformational angles using a bivariate von Mises distribution and Bayesian nonparametrics. *J. Am. Stat. Soc.* **104**, 586-596.

4. Rotkiewicz,P. & Skolnick,J. (2008). Fast procedure for reconstruction of ful-atom protein models from reduced representations. *J. Comp. Chem.* **29**, 1460-1465.
5. Voronai,G.F. (1908). Nouveles applications des paramètres continus a la théorie des formes quadratiques, *J. Reine Angew. Math.* **134**, 198-287.
6. Joo,H., Qu,X., Swanson,R., McCallum,C.M., & Tsai,J. (2010). Fine grained sampling of residue characteristics using molecular dynamics simulation. *Comput. Biol. Chem.* **34**, 172-183.

4_BODY_POTENTIALS

CASP9: Four Body Potentials for the Prediction of Protein Structure

Sumudu P. Leelananda[1], Pawel Gniewek[1, 2], Andrzej Kloczkowski[1*]
[1]*Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, Ames, IA, USA*
[2]*Department of Chemistry, University of Warsaw, Warsaw Poland*
*kloczkow@iastate.edu

Multi-body potentials have been of much interest recently because they take into account three dimensional interactions related to residue packing and capture the cooperativity of these interactions in protein structures. We combined long range multi-body potentials and short range potentials to improve recognition of native structure among misfolded decoys. We optimized the weights for four-body non-sequential, four-body sequential and short range potential in order to obtain optimal model ranking.

Our optimized multi-body potentials outperform all other contact potentials in the recognition of the native structure among decoy sets, both for models from homology-based modeling and from template-free modeling in CASP8 decoy sets. We have compared the results obtained for this optimized potential, with those from the DFIRE potential, which takes into account atomic level information of proteins empirically. We find that for all proteins larger than 80 amino acids optimized coarse-grained potentials yield results comparable to those obtained with the atomic DFIRE potential.

One of the most widely used two-body potentials in the assessment of protein models is the Miyazawa-Jernigan potential [1]. Betancourt and Thirumalai suggested that pair-wise potentials are not likely to be sufficient for threading applications [2]. The alternative multi-body potentials in principal are able to take account of more complex three dimensional interactions, revealing the effects of dense residue packing. Importantly they can capture the strong cooperativity operative within protein structures. Three-body potentials were proposed and developed by Munson and Singh [3] and they all showed improvements over two-body potentials. Four-body potentials were first derived in the context of Delaunay tessellation by Krishnamoorthy and Tropsha [4] and they demonstrated that these potentials also perform better than two-body potentials.

The four-body contact potentials developed by our group[5] incorporated sequence information and considered in detail the interactions between backbones and side chains through a simple geometric construction. We also developed them to distinguish between different levels of solvent accessibility of the residues.

Further we have improved the performance of the four-body contact potentials by combining the four-body sequential ([5]) with the four-body non-sequential potentials [6] and with short range potential and have used this optimized potential in the identification protein native structure.

Methods

We obtained predictions from several servers that performed well in CASP8. These servers include Zhang, Baker, Raptor, HHPred, Tasser, Pcons and SAM servers. All the predictions from each of these servers were taken which amounts to 30 structure predictions in total. Optimized four body potentials were applied to each of these structures and the minimum energy given structure was identified

as the best fit to the native. In each case, the identified structure was visualized using software to make sure it is a reasonable model.

Availability

Four-body sequential, four-body non-sequential and short ranged potentials are freely available in our Potentials 'R' Us web server ([6]).

1. Miyazawa, S. & Jernigan, R. L. (1996). Residue - Residue Potentials with a Favorable Contact Pair Term and an Unfavorable High Packing Density Term, for Simulation and Threading. *Journal of molecular biology* **256**, 623-644.
2. Betancourt, M. & Thirumalai, D. (1999). Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein science* **8**, 361.
3. Munson, P. & Singh, R. K. (1997). Statistical significance of hierarchical multi-body potentials based on Delaunay tessellation and their application in sequence-structure alignment. *Protein science* **6**, 1467.
4. Li, X. & Liang, J. (2005). Geometric cooperativity and anticooperativity of three-body interactions in native proteins. *Proteins* **60**, 46.
5. Krishnamoorthy, B. & Tropsha, A. (2003). Development of a four-body statistical pseudo-potential to discriminate native from non-native protein conformations. *Bioinformatics* **19**, 1540-1548.
6. Feng, Y., Kloczkowski, A. & Jernigan, R. L. (2007). Four-body contact potentials derived from two protein datasets to discriminate native structures from decoys. *Proteins* **68**, 57-66.
7. Feng, Y., Kloczkowski, A. & Jernigan, R. (2010). Potentials 'R'Us web-server for protein energy estimations with coarse-grained knowledge-based potentials. *BMC Bioinformatics* **11**, 92.

ALAdeGAP

Improvement of the Quality of Model Structures by Improving the Template-Target Alignments

K. Yura[1], A. Hijikata[2] and M. Go[3,4,5]
[1] Computational Biology, Graduate School of Humanities and Sciences, Ochanomizu University, 2-1-1 Otsuka, Bunkyo, Tokyo 112-8610, Japan, [2] Laboratory for Immunogenomics, RIKEN Research Center for Allergy and Immunology, 1-7-22 Suehiro, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan, [3] Department of Bioscience, Faculty of Bioscience, Nagahama Institute of Bio-Science and Technology, 1266 Tamura, Nagahama, Shiga 526-0829, Japan, [4] Tokyo Medical and Dental University, 1-5-45 Yushima, Bunkyo, Tokyo 113-8510, Japan, [5] Research Organization of Information and Systems, 4-3-13, Toranomon, Minato, Tokyo 105-0001, Japan
yura.kei@ocha.ac.jp

The quality of alignment for comparative modeling still constitutes a major bottleneck to obtain high quality in computationally modeled protein three-dimensional (3D) structures[1,2]. We improved the quality of the alignment by adjusting the method to introduce gaps in the alignment for comparative modeling and assessed the efficacy of the new alignment method by building a protein model structures based on a conventional alignment and the new alignment.

Methods

We revisited the correlation between protein 3D structure and the gap location in a large protein 3D structure data set, and found that the frequency of the gap location was approximated with the exponential function of the solvent accessibility of the inserted residues. The relationship was previously considered as linear based on a small data set. We introduced this newly found relationship to gap penalty calculation of the alignment between template and target sequences. In the template and target sequence alignment, at least one of the sequences has a known 3D structure by definition, and the 3D structure information can be used to calculate the solvent accessibility of each residue in the alignment. Gap penalties in the alignment were then calculated based on these solvent accessibility values.

Results

Only by modifying the gap penalty calculation method, the sequence alignment much closer to the structural alignment was obtained. The quality of the alignment was substantially improved on a pair of sequences with identity in a twilight zone, approximately around 20 to 40%. In a benchmark test, we found that the protein model structures built on a conventional alignment and the new alignment were different at the location of loops, and that the structures built on the new alignment were much closer to the structures determined by X-ray crystallography[3].

Availability

The method is implemented in a computer program ALAdeGAP (ALignment with Accessibility dependent GAp Penalty) and is available at http://cib.cf.ocha.ac.jp/target_protein/.

1. Kopp,J., Bordoli,L., Battey,J.N.D., Kiefer,F., Schwede,T. (2007). Assessment of CASP7 predictions for template-based modeling targets. *Proteins* **69** (Suppl 8), 38-56.
2. Keedy,D.A., Williams,C.J. Headd,J.J., Arenadall,W.B.III, Chen,V.B., Kapral, G.J., Gillespie,R.A., Block,J.N., Zemla,A., Richardson,D.C., Richardson,J.S. (2009). The other 90% of the protein:

assessment beyond the Cαs for CASP8 template-based and high-accuracy models. *Proteins* **77** (Suppl 9), 29-49.

3. Hijikata,A., Yura,K., Noguti,T., Go,M. Revisiting gap locations in amino acid sequence alignment and a proposal for a method to improve them by introducing solvent accessibility. *submitted*.

AOBA

Quality Assessment by Structural Consensus and Statistical Scoring Functions and Modeling by Hybridization of Server Models

M. Shirota[1] and K. Kinoshita[2]
[1] – *Graduate School of Information Science, Tohoku University,* [2] – *Institute for Bioinformatics Research and Development, JST*
mshirota@hgc.jp

AOBA team participated in CASP9 as a human group for Tertiary Structure (TS) and Quality Assessment (QA) categories.

Methods

Our QA method attempts to evaluate both the quality of each server model and that of each residue in the models. We assumed that the average structure of the best server models would approximate the native structure and defined the quality of the models and residues with reference to this average structure. The server models for each target were first ranked using structural consensus score and two statistical scoring functions. We evaluated the TM-score[1] of each server model to all the remaining models and used the average TM-score as the structural consensus score. We used Verify3D[2] and Stability[3] function as statistical scoring functions. The average TM-score and the per-residue Verify3D and Stability scores were added with equal weights. We chose top 16 models by this combined score and generated their average structure by iterating the superposition and averaging of two structures. We defined the quality of each server model by its TM-score from this average structure. The quality of each residue was defined by the distance between its CA atom and the corresponding atom in the average structure after structural superposition. The lower limit of the quality of each residue was set the half of the average deviation of the corresponding residues in the top 16 ranked models, whereas the upper limit was set 9.9.

AOBA TS method attempts to hybrid two server models to generate a refined structure. Each server model was selected as a seed with a probability proportional to $\exp\{\alpha * \mathrm{score}\}$, where score is the combination score used in our QA method and α is a constant. We selected 1,000 to 3,000 different pairs of server models. The two selected server models of each pair were superposed and the distance of the corresponding residues in the two models were measured. The residues in the protein were classified as either the core residues (the distance within 5 Å) or the loop residues (the distance over 5 Å). We generated up to 10 alignments of the query amino acid sequence and the two server models. In these alignments, the core residues of the query sequence were aligned with both of the two models, whereas each stretch of the loop residues were aligned with only one of the models at random. We generated structures from these alignments using MODELLER program[4]. The generated models from all the alignments of all the seed pairs were ranked using Stability function and the best 5 models were submitted.

1. Zhang,Y., & Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins.* **57**, 702-710.
2. Bowie,J.U., Juthy,R., & Eisenberg,D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253**, 164-170.

3. Ota,M. & Nishikawa,K. (1997) Assessment of pseudo-energy potentials by the best-five test: a new use of the three-dimensional profiles of protein. *Protein Eng.* **10**, 339-351.
4. Sali,A. & Blundel,T.L. Comparative protein modeling by satisfaction of spatial restraints. (1993) *J. Mol. Biol.* **234**, 779-815.

Atome2_CBS

Jean-Luc Pons, Gilles Labesse
J-L;Centre de Biochimie Structurale / CBS CNRS UMR 5048 - UM 1 - UM 2 - INSERM UMR 554 - 29 rue de Navacelles. 34090 MONTPELLIER France

@TOME 2.0 (1) is new web pipeline dedicated to protein structure modeling and small ligand docking based on comparative analyses. @TOME 2.0 allows fold recognition, template selection, structural alignment editing, structure comparisons, 3D-model building and evaluation. These tasks are routinely used in sequence analyses for structure prediction. In our pipeline the necessary bioinformatic tools were efficiently interconnected in an original manner to accelerate all the processes. Furthermore, we have also connected comparative docking of small ligands that is performed using protein–protein superposition. The input is a simple protein sequence in one-letter code with no comment. The resulting 3D model, protein–ligand complexes and structural alignments can be visualized through dedicated Web interfaces or can be downloaded for further studies.

The sequences submitted to CASP9 were automaticaly treated as follows:

The best structural alignment (SA) are extracted from each fold recognition software result: Psiblast (2), Hhsearch (3), Fugue (4), Sp3 (5). For each SA, a 3D common core is generated by TITO (6).

On the overall results, the 20 best SA are selected according a global score (@TOME-2 Score) based on a set of quality descriptors: Fold recognition tools score, sequence identity between query/template, quality of alignment (T-coffee, 7), compatibility between amino acid sequence and 3D template (TITO), Verify3D (8) & QMean (9) evaluation scores of model after sides chains calculation with Scwrl software (10). Structural clusters are calculated (Maxcluster, 11) and all the SA outside the main cluster are rejected.

In a second step, eight multi-template models were computed by MODELLER 9.0 (12). For each models to construct, 4 templates have been selected according the best scores from @TOME-2, Verify3D, TITO and Qmean. For each group of template, the MODELLER model is calculated with and without a prior step of structural realignment via Matt (13). Among the 8 models obtained, the 5 best QMean score have been proposed to CASP9.

Moreover, the comparative docking of biologically relevant ligands from PROCOGNATE (14) have been used for automatically detect active sites.

Availability: http://atome.cbs.cnrs.fr/

Reference:

1. Pons,JL. & Labesse,G. (2009). @TOME-2: a new pipeline for comparative modeling of protein-ligand complexes. Nucleic Acids Research, Web Server Issue 2009 - doi: 10.1093/nar/gkp368.
2. Altschul et al (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25(17): 33100-3402
3. Soding,J. (2005). Protein homology detection by HMM-HMM comparison. Bioinformatics, Bioinformatics. 21(7): 951-60.
4. Shi et al (2001). FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure- dependent gap penalties. J. Mol. Biol., 310, 243-257.

5. Zhou,H. & Zhou,Y. (2005). Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments, PROTEINS: Structure, Function, and Bioinformatics 58:321–328
6. Labesse,G. and Mornon,J-P. (1998). Incremental threading optimization (TITO) to help alignment and modelling of remote homologues. Bioinformatics, 14, 206-350
7. Notredame,C. Higgins,DG. Heringa,J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol ,302(1):205-17.
8. Eisenberg,D. Lüthy,R. Bowie,JU (1997). VERIFY3D: assessment of protein models with three-dimensional profiles. Methods Enzymol. 277:396-404.
9. Benkert,P. Tosatto,S.C.E. and Schomburg,D. (2008). "QMEAN: A comprehensive scoring function for model quality assessment." Proteins: Structure, Function, and Bioinformatics, 71(1):261-277.
10. Canutescu,A. Shelenkov,A. and Dunbrack,R. L. (2003). A graph theory algorithm for protein side-chain prediction. Protein Science 12, 2001-2014.
11. Ortiz,A.R., Strauss,C.E. and Olmea,O. (2002). MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. Protein Sci, 11, 2606-21.
12. Eswar,N. Eramian,D. Webb,B. Shen,M. Sali,A. (2006). Protein Structure Modeling With MODELLER. Methods in Molecular Biology, 2008, Volume 426, 1, 145-159.
13. Menke,M. Berger,B. Cowen,L. (2007). "Matt: Local Flexibility Aids Protein Multiple Structure Alignment", PLoS Comput Biol 4(1): e10. doi:10.1371/journal.pcbi.0040010
14. Bashton,M et al. (2008). PROCOGNATE: a cognate ligand domain mapping for enzymes. Nucleic Acids Research 36: D618-D622.

BAKER

Modeling of Protein Structures Using Rosetta in CASP9

J. Thompson[1], T.J. Brunette[1], D.E. Kim[1], F. Khatib[1], D. Gront[1], F. DiMaio[1], R. Wang[1], R. Vernon[1], B. Kim[2], J. Pei[2], S. Cooper[1], M. Tyka[1] and D. Baker.[1]
[1]- University of Washington, [2] - University of Texas Southwestern Medical Center
dabaker@u.washington.edu

The primary change to our structure prediction protocol in CASP9 is an iterative modeling procedure that attempts to improve on our automated comparative modeling protocol by performing restraint-based minimization of the Rosetta full-atom energy[5]. Other new features include the use of spatial restraints derived from template structures, comparative modeling using symmetry inferred from templates[2], and use of Foldit[4] to examine comparative modeling problems in real-time.

Methods

We used previously described domain-parsing and disorder prediction algorithms to parse sequences into domains[1]. For each domain, we attempted to identify homologous template structures using sequence-based methods[3]. When obvious templates for modeling were available, we built comparative models using the standard Rosetta rebuild and refine protocol[3]. In cases with no clear similarity to known structures we assembled models using our standard free modeling protocol[3].

We made comparative models of protein structures by feeding the above alignments into the standard Rosetta rebuild and refine protocol. Model refinement used spatial restraints derived from template structures to prevent excessive divergence from the templates, and refinement was primarily guided by the Rosetta full-atom energy. After generating the first round of models using an automated protocol, we examined alignments, template structures and low-energy models using Foldit[4]. In some cases we modeled proteins using symmetry inferred from template structures[2]. For small proteins with clear homology to known structures, we used an iterative protocol developed for NMR structure determination[5]. The iterative protocol started with models from the standard comparative modeling procedure, and distance restraints from templates replaced the NMR restraints. Comparative models were selected based on all-atom Rosetta energy and visual inspection.

Our free modeling protocol builds models from extended protein chains using fragments of known protein structures and a low-resolution representation of protein side-chains[3]. For proteins with beta strands, we constructed high contact-order structures by explicitly enforcing beta-strand pairings and by energetically penalizing low contact-order structures. Following fragment assembly, structures were refined with the standard Rosetta refinement procedure, which explicitly represents all heavy and hydrogen atoms while minimizing the Rosetta all-atom energy function. Low-energy models were clustered, and models were selected by a combination of visual inspection and Rosetta full-atom energy.

Results

Using the iterative protocol we produced several models that clearly improved on the starting templates, including T0520 and T0580, both close comparative modeling targets. Human inspection and modification of alignments proved successful for several targets, such as T0556, T0569 and T0614. In free modeling, we had several notable successes, including T0581 and T0624. Modeling failures resulted from selection of incorrect alignments in building models, inclusion of inaccurate templates in building spatial restraints, over-ordering of disordered regions, and incomplete sampling of the rugged Rosetta full-atom energy landscape. We are currently working to develop more effective search measures that

simultaneously sample both the Rosetta full-atom energy landscape and the input parameters that guide conformational search.

Availability

The automated portion of the methods described herein are available from the Rosetta Commons, at http://www.rosettacommons.org.

1. Kim DE, Chivian D, Malmström L, Baker D. (2005). "Automated prediction of domain boundaries in CASP6 targets using Ginzu and RosettaDOM." *Proteins*. **61,** 193-200.
2. André I, Bradley P, Wang C, Baker D. (2007). "Prediction of the structure of symmetrical protein assemblies." *PNAS*. **104**:17656-17661.
3. Raman S, Vernon R, Thompson J, Tyka M, Sadreyev R, Pei J, Kim D, Kellogg E, DiMaio F, Lange O, Kinch L, Sheffler W, Kim BH, Das R, Grishin NV, Baker D. "Structure prediction for CASP8 with all-atom refinement using Rosetta." *Proteins*. **77**, 89-99.
4. Cooper S, Khatib F, Treuille A, Barbero J, Lee J, Beenen M, Leaver-Fay A, Baker D, Popovic Z, Players F. (2010). "Predicting protein structures with a multiplayer online game." *Nature*. **466**, 756-760.
5. Raman S, Lange OF, Rossi P, Tyka M, Wang X, Aramini J, Liu G, Ramelot TA, Eletsky A, Szyperski T, Kennedy MA, Prestegard J, Montelione GT, Baker D. (2010), "NMR structure determination for larger proteins using backbone-only data." *Science*. **19**, 1014-1018.

Baltymus

Quality assesment of single protein structure models using geometrical and statistical techniques

K. Olechnovič, M. Margelevičius and Č. Venclovas
Institute of Biotechnolgy, Graičiūno 8, LT-02241 Vilnius, Lithuania
kliment@ibt.lt

Baltymus evaluates individual protein structure models based on geometrical and statistical considerations. In CASP9 "Baltymus" was used to produce 5 quality assesment predictions for each target with the intention of testing different combinations of computed quality scores.

Methods

Given a model, we represent its atoms as spheres of Van der Waals radii. Accordingly, the model 3D structure is a geometric object formed by such spheres. We subdivide this object into quadruples of spheres, making sure that no tangent sphere of any quadruple overlaps any atom sphere. The subdivision is known as the Apollonius graph (Emiris and Karavelas, 2006) and is similar to the Delaunay triangulation, except that spheres are used instead of points and tangent spheres are used instead of circumspheres. We use the Apollonius graph to identify the cavities surrounding each atom of the model and to calculate the volume of the cavities. For each amino acid type we had defined a potential function based on the cavities statistics obtained from the PISCES culled PDB set of protein structures with the percentage identity cutoff of 20% and the resolution cutoff of 1.8 angstroms. This function assigns a quality score to the volume of the cavities surrounding each residue. The model score is a sum of the residues scores. We also compute another quality estimate term using a version of knowledge-based statistical pairwise potentials (Sippl, 1993) for C, CA, CB, N and O atoms. Then we normalize the two obtained scores by the number of model residues and convert the normalized scores into p-values. The final quality estimate is produced from the weighted combination of p-values (Theiler, 1996).

Availability

Our method is still in early stages of development and a user-friendly software package dedicated to protein structure models quality assessment is not yet available. However, it is possible to explore the geometrical aspects of the method using our recently developed interactive tool, Voroprot, which can be downloaded from http://www.ibt.lt/bioinformatics/software/voroprot.html.

1. Emiris, I.Z. and Karavelas, M.I. (2006). The predicates of the Apollonius diagram: Algorithmic analysis and implementation, Comput. Geom.-Theory Appl., 33: 18-57.
2. Sippl, M. J. (1993). Recognition of errors in three-dimensional structures of proteins. Proteins, 17:355-62.
3. Theiler, J. (1996), Combining Statistical Tests By Multiplying p-values.
4. Wang, G. and Dunbrack, R. L. (2003). PISCES: a protein sequence culling server. Bioinformatics, 19:1589-1591.

BHAGEERATH

Bhageerath: an energy based web-enabled computer software suite for predicting the tertiary structures of soluble proteins

Priyanka Dhingra, Bharat Lakhani, Shashank Shekhar, Avinash Mishra, Ashutosh Shandilya and B.Jayaram*
Department of Chemistry and Super Computing Facility for Bioinformatics and Computational Biology, Indian Institute of Technology, Hauz Khas, New Delhi-110016, India
bjayaram@chemistry.iitd.ac.in; priyanka@scfbio-iitd.res.in

Bhageerath(http://www.scfbio-iitd.res.in/bhageerath/index.jsp) is an energy based computer software suite for predicting the tertiary structures of soluble proteins. The protocol is initiated with a prediction of the secondary structures from the input amino acid sequence, candidate structures are generated by an extensive sampling of the conformational space of the loop regions, improbable structures are filtered out with some biophysical filters, the resultant structures are energy ranked and five best structures-energy wise are selected, which are further refined using explicit solvent molecular dynamics simulations.

Methods

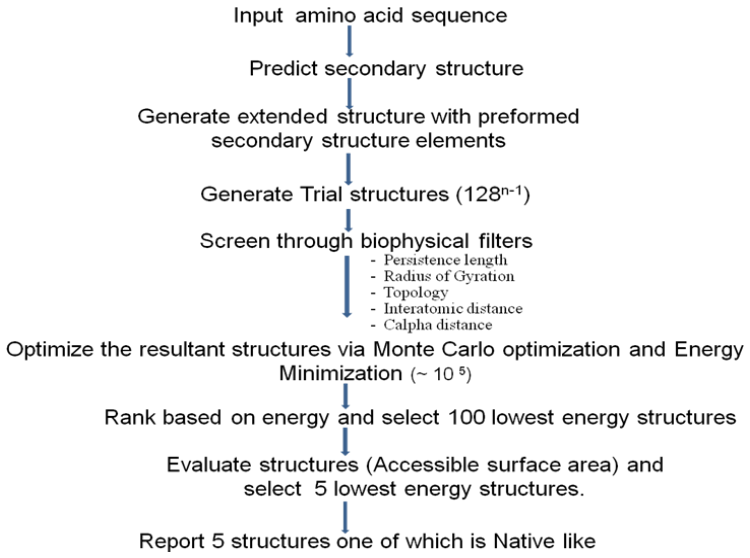


Figure 1: Bhageerath's pathway for predicting structures of small globular proteins with less than eight secondary structural elements.

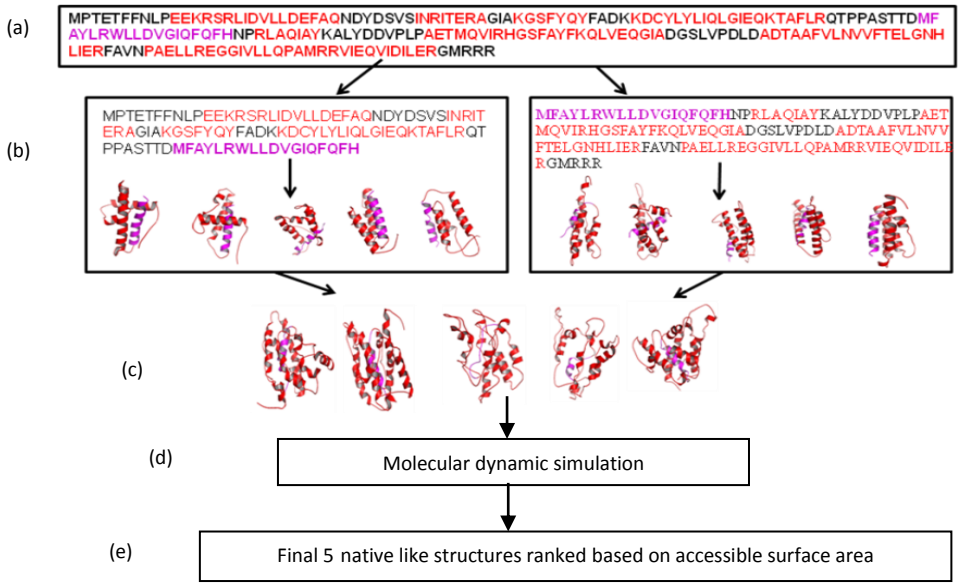


Figure 2: Bhageerath's pathway for predicting tertiary structures of large proteins. (a) Amino acid sequence is taken as input to the protocol. (b) The input sequence is fragmented into individual fragments with common secondary structural termini and each fragment is processed separately. (c) The predicted 5 structures from individual fragments are patched and energy minimized to rank top 5 lowest energy structures. (d) Selected top 5 structures are submitted for molecular dynamic simulations. (e) Final 5 candidate structures for the native are ranked based on accessible surface area.

1. Narang,P., Bhushan,K., Bose,S. & Jayaram,B. (2005). A computational pathway for bracketing native-like structures for small alpha helical globular proteins. *Phys. Chem. Chem. Phys.* **7**, 2364-2375.
2. Narang,P., Bhushan,K., Bose,S., & Jayaram,B. (2006). Protein structure evaluation using an all-atom energy based empirical scoring function. *J. Biomol. Str. Dyn.* **23**, 385-406.
3. Jayaram,B., Bhushan,K., Shenoy,S.R., Narang,P., Bose,S., Agrawal,P., Sahu,D., Pandey,V.S. Bhageerath : An Energy Based Web Enabled Computer Software Suite for Limiting the Search Space of Tertiary Structures of Small Globular Proteins. *Nucleic Acids Res.* **34**, 6195-6204.

BHAGEERATH_SCFBIO

Bhageerath-H: An ab-initio, homology combined hybrid model for protein tertiary structure prediction

Pallavi Mohanty, Bharat Lakhani, Priyanka Dhingra, Avinash Mishra, Ashutosh Shandilya, B.Jayaram*
Department of Chemistry and Super Computing Facility for Bioinformatics and Computational Biology, Indian Institute of Technology, Hauz Khas, New Delhi-110016, India
bjayaram@chemistry.iitd.ac.in; priyanka@scfbio-iitd.res.in; pallavi@scfbio-iitd.res.in

Homology modeling tools work exceptionally well if a query sequence finds a similarity against a reference pdb in RCSB (www.rcsb.org). As the similarity vanishes, alternative methods become imperative. We describe here a hybrid method which is a combination of BHAGEERATH, an *ab-initio* method for protein structure prediction and homology modeling tools like Phyre and Modeller.

Given a target protein sequence, we initially look for a homologous protein(s) with known 3D structures using PSI-BLAST or BLASTP. The template structures are used to build the models using publicly available homology modeling software's. However, missing residue regions, which do not show significant sequence similarity, are submitted to BHAGEERATH. BHAGEERATH [1-3] is an all atom energy based software to predict tertiary structures of soluble proteins. It predicts five candidate structures and each of the predicted structures is patched with the homology modeled structure to put together candidates for the whole tertiary structure of the protein. The five predicted structures are energy minimized and further refined using explicit solvent molecular dynamics simulations. The methodology will be available very soon as a web server christened BHAGEERATH-H.

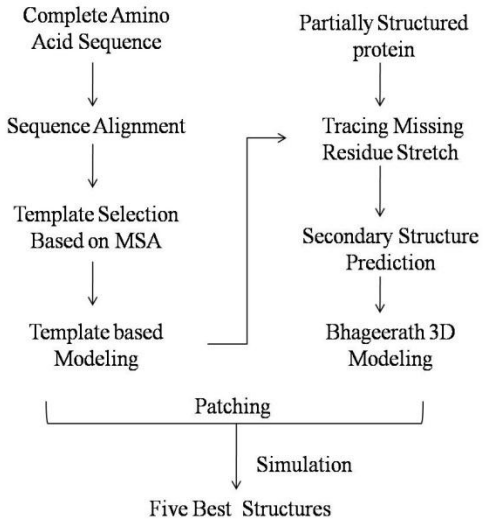


Figure 1: BHAGEERATH-H pathway for protein tertiary structure prediction.

Preliminary assessment of this methodology on CASP9 targets indicates that in at least half of the targets under human group whose native structures are released, a root mean square deviation of < 7 Å is realized vis-a-vis native. Several improvements are envisioned to the protocol for better results.

1. Narang,P., Bhushan,K., Bose,S. & Jayaram,B. (2005). A computational pathway for bracketing native-like structures for small alpha helical globular proteins. Phys. Chem. Chem. Phys. 7, 2364-2375.
2. Narang,P., Bhushan,K., Bose,S., & Jayaram,B. (2006). Protein structure evaluation using an all-atom energy based empirical scoring function. *J. Biomol. Str. Dyn.* **23**, 385-406.
3. Jayaram,B., Bhushan,K., Shenoy,S.R., Narang,P., Bose,S., Agrawal,P., Sahu,D., Pandey,V.S. (2006). Bhageerath : An Energy Based Web Enabled Computer Software Suite for Limiting the Search Space of Tertiary Structures of Small Globular Proteins. *Nucleic Acids Res.* **34**, 6195-6204.

4. Kelley, L.A., & Sternberg, M.J.E. (2009). Protein structure prediction on the web: a case study using the Phyre server. *Nature Protocols*. **4**, 363-371.
5. http://salilab.org.modeller/

Bilab-ENABLE
Bilab-solo
Bilab

Tertiary Structure Prediction by Combination of Fold Recognition, Realignment, Fragment Assembly and Consensus-based Model Quality Prediction

S. Nakamura[1], M. Morita[2] and M. Kakuta[1]
[1] - *Department of Biotechnology, The University of Tokyo,*
[2] – *Laboratory of Bioinformatics, National Institute of Biomedical Innovation*
shugo@bi.a.u-tokyo.ac.jp

We have attended to CASP9 as three prediction groups. Bilab-ENABLE is a full-automated prediction server. Bilab-solo is a human prediction group without any information of server models except for models generated by our group. Human experts added some models to models from Bilab-ENABLE and pickup five best models among them. Bilab is also a human prediction group. It used a semi-automated procedure based on fragment assembly method.

Methods

The following is the overview of the procedure of our ENABLE server: 1) Templates for a target were first searched and template-target alignments were generated using PDB-BLAST, FUGUE[1], and HHpred[2] combined with T-COFFEE[3]. 2) Another alignments were generated by our realignment technique named REALIZE. REALIZE is a pairwise sequence alignment tool based on profile-profile comparison utilizing structure-dependent gap penalties and predicted secondary structures. Structure-dependent gap penalties were calculated according to residue environments defined by secondary structure, atom depth, and hydrogen-bonding pattern of the template structures. 3) First set of models were generated by using MODELLER[4] from a variety of template-alignment combinations generated on step 2. 4) Starting from the structures generated on step 3, refinement procedure based on fragment assembly called IDDD/ABLE[5] developed in our laboratory was executed and added to the model structure set. Target function including burial of hydrophobic residues, contacts between residues, average distance between hydrophobic residues, hydrogen bonds between mainchains, and exclusive volume to avoid overlap of residues was minimized by simulated annealing with 5000-20000 steps. For each refinement, about 5000 models were generated, clustering was applied and centers of five largest clusters were picked up. The size of model set was about 3000 after this step. 5) Top 500 models were selected according to Verify3D[6] scores. Qualities of the models were then assessed by our developed QA predictor based on consensus method and five best models were selected for submission. 6) Predicted quaternary structures were generated considering quaternary structures of homologs on PISA server[7]. Ligand-binding sites were also predicted making use of ligand coordinates in homologs. When homologs including ligands were not found, energy-based ab initio method was applied to model structures.

During prediction by Bilab-solo group, we had not used any server predictions at all except for models by our ENABLE server. For most cases, sequence alignments constructed by ENABLE were corrected manually, and then they were served to MODELLER to construct models. Models from 2nd round of ENABLE, generated by IDDD/ABLE using fragments with 21 amino acids from models of 1st round of ENABLE and those with 9 amino acids from non-redundant protein structure database starting from extended structures, were also added to the model set. Model qualities and ranks were estimated by manual inspection.

Group Bilab is similar to 2nd round of ENABLE, except that it used top 300 models of Verify3D scores from models by 1st round of ENABLE server and server models by other servers on CASP9 web site.

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. Söding J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**, 951-960.
3. Notredame,C., Higgins,D.G. & Heringa,J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205-17.
4. Sali,A. & Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779-815.
5. Ishida,T., Nishimura,T., Nozaki,M., Inoue,T., Terada,T., Nakamura,S. & Shimizu,K. (2003) Development of an ab initio protein structure prediction system ABLE. *Genome Inform.* **14**, 228-237.
6. Luthy,R., Bowie,J.U. & Eisenberg,D. (1992) Assessment of protein models with three-dimensional profiles. *Nature* **356**, 83-85.
7. Krissinel,E. & Henrick,K. (2007). Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372**, 774-797.

BIO_ICM

Protein structure modeling using 3D-Jury and pyROSETTA

K. Steczkiewicz[1,3*], M. Lazniewski[1,2] and T. Wlodarski[1,3]
[1] - *Interdisciplinary Centre for Mathematical and Computational Modelling*
[2] - *Department of Physical Chemistry, Faculty of Pharmacy, Medical University of Warsaw Poland*
[3] - *The Inter-Faculty Individual Studies in Natural Science and Mathematics, University of Warsaw, Poland*
*ksteczk@icm.edu.pl

In CASP9 experiment we modeled 57 targets using consensus fold recognition 3D-Jury server supported by pyROSETTA *de novo* simulation and molecular dynamics for low homology regions and overall structure optimization

Comparative modeling approach comprises of four major steps: **1)** template identification; **2)** target-to-template structure-based alignment; **3)** structure modeling and optimization; and **4)** model verification.

1) Template identification is an essential step validating homology modeling method applied for particular target. It strongly relies on sequence comparison method's quality: its sensitivity and selectivity. We used consensus fold recognition approach – 3D-Jury server[1] in order to choose known protein structures as potential templates for following modeling steps. **2)** In order to build probabilistic model for every target we collected sequences of its closest homologs with PSI-Blast[2] exhaustive searches until profile convergence (e-value threshold 0.005). Then, PCMA[3] was used to calculate multiple sequence alignment of collected sequences. For every sequence of the alignment secondary structure was predicted with PSI-PRED[4] which allowed us to study conservation of secondary structure elements across the family. Simultaneously, structural alignment was generated for previously identified templates. Finally, we acquired two alignments – one for target's family (extended with secondary structure profile) and second for superimposed templates sequences. Therefore, we were able to prepare high-accuracy sequence-to-structure alignment respecting conserved hydrophobic residue patches and secondary structure elements arrangement. **3)** Obtained target-to-templates alignment was fed to modeling software, MODELLER[5]. When possible, we used multiple template structures (even less similar to the target than the best scoring template but still preserving structural core of the fold) in order to allow modeling program for major local adjustments often not allowed by standard optimization methods. Regions with no alignment to known structures were modeled *de novo* with ROBETTA[6] server and molecular dynamics implemented in Tripos SYBYL software. Modeled sidechains were further optimized with SCRWL[7] and finally the model was optimized with pyROSETTA[8] Monte Carlo relaxation routine in order to reduce overall structure energy. **4)** Afterwards, obtained model was subjected to servers used for X-ray or NMR structures rigorous evaluation: PROSA[9] and MolProbity[10] to identify regions that violate general laws ruling protein structure stability. Poorer regions of the model were manually investigated in detail and eventually subjected for local remodeling.

1. Ginalski, K., Elofsson, A., Fischer, D. and Rychlewski, L. (2003) 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics*, **19**, 1015-1018.
2. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389-3402.

3. Pei, J., Sadreyev, R. and Grishin, N.V. (2003) PCMA: fast and accurate multiple sequence alignment based on profile consistency. *Bioinformatics*, **19**, 427-428.
4. Jones, D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*, **292**, 195-202.
5. Eswar, N., Webb, B., Marti-Renom, M.A., Madhusudhan, M.S., Eramian, D., Shen, M.Y., Pieper, U. and Sali, A. (2006) Comparative protein structure modeling using Modeller. *Curr Protoc Bioinformatics*, **Chapter 5**, Unit 5 6.
6. Kim, D.E., Chivian, D. and Baker, D. (2004) Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res*, **32**, W526-531.
7. Wang, Q., Canutescu, A.A. and Dunbrack, R.L., Jr. (2008) SCWRL and MolIDE: computer programs for side-chain conformation prediction and homology modeling. *Nat Protoc*, **3**, 1832-1847.
8. Chaudhury, S., Lyskov, S. and Gray, J.J. (2010) PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics*, **26**, 689-691.
9. Wiederstein, M. and Sippl, M.J. (2007) ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res*, **35**, W407-410.
10. Chen, V.B., Arendall, W.B., 3rd, Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J., Murray, L.W., Richardson, J.S. and Richardson, D.C. (2010) MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr*, **66**, 12-21.

BIOMINE

Prediction of Disorder Regions by Multilayer Information Fusion

M.J. Mizianty[1], W. Stach[1], K. Chen[1], K.D. Kedarisetti[1], F.M. Disfani[1] and L. Kurgan[1]
[1] *Department of Electrical and Computer Engineering, University of Alberta, Canada*
lkurgan@ece.ualberta.ca

We redesigned our recent method, MFDp[1], which generates predictions characterized by high MCC values, to build 4 predictors that we utilized in the CASP9 experiment. The original MFDp predictor was designed using disordered regions for which annotations were extracted from both PDB and DisProt[2], and using strict limits on sequence similarities between training and test datasets. We designed our CASP9 predictors using different disorder annotations, i.e., we considered solutions trained on both disorder annotations and solutions trained on only the PDB-based annotation, and we included a larger number of training chains that were not bound by the similarity limits.

Methods

The design of input features, features selection and training procedure were adopted from the original paper[1]. MFDp is a meta approach which combines disorder predictions from three complementary predictors: machine learning-based DISOPRED2[3], structure prediction-based DISOclust[4], and residue propensity-based IUPred[5]. Unlike a number of other consensus-based disorder predictors, MFDp also includes other input information sources, which include evolutionary profiles (in the form of PSSM), and predicted secondary structure, solvent accessibility, residues flexibility (B-factor), back-bone dihedral torsion angles, and globular domains. The input features include raw values, as well as aggregated predictions including maximal, minimal and average values over the window around a predicted residue. We also annotate local predicted secondary structures conformations. Three subsets of selected features are feed into three corresponding Support Vector Machines specialized for the prediction of short, long and generic (all) disordered regions. The final predicted probability of disorder is computed as the maximal value over the three SVM-based outputs.

The main differences between the MFDp and the methods used during the CASP9 are the training datasets used to build the SVM models, different composition of the SVM classifiers used to generate the final predictions, and use of two post-prediction filters. Our group registered four disorder predictors: biomine_DR_mixed, biomine_DR_mixed_c, biomine_DR_pdb, and biomine_DR_pdb_c. The two "mixed" predictors were trained on a subset of the MxD dataset[1] (we used only proteins from the PDB and fully disordered proteins from the DisProt) and their predictions were computed as the average over two SVMs for the short and the long disordered regions. The "pdb" predictors were trained on new dataset created using PDB depositions and the predictions were generated using the maximal value of predicted probability from two SVMs for the long and all disorder regions. The methods with "_c" suffix were optimized for the MCC values, whereas the methods without the suffix were optimized for the S_w measure.

We also applied two post-prediction filters which work at the sequence level. First, instead of reporting raw predicted probability values for each residue we aggregate probabilities using the mean value over 5-residues window. Second, we remove short, up to 3 residues, disorder/ordered segments.

Results

We evaluated our methods on two datasets, CASP8 and the dataset used to train the "pdb" methods. The results are summarized in the table below. The "pdb" methods provide better predictions,

which is not surprising, since they were built using PDB annotations, which dominate both of the datasets.

Method	CASP8 dataset			PDB-derived dataset		
	AUC	MCC	Sw	AUC	MCC	Sw
biomine_DR_mixed	0.894	0.635	0.629	0.894	0.415	0.613
biomine_DR_mixed_c	0.894	0.635	0.469	0.894	0.425	0.517
biomine_DR_pdb	0.915	0.568	0.685	0.915	0.450	0.666
biomine_DR_pdb_c	0.915	0.652	0.564	0.915	0.522	0.585

Availability

The original MFDp predictor, which is the precursor for our four registered predictors, is freely available on-line as a web server and a standalone application at http://biomine.ece.ualberta.ca/MFDp.html.

1. Mizianty,M.J., et al. (2010). Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources, *Bioinformatics* **26**: i489-i496.
2. Sickmeier,M., et al. (2007). DisProt: the database of disordered proteins. *Nucleic Acids Res.* **35**:D786-93.
3. Ward,J.J., et al. (2004).The DISOPRED server for the prediction of protein disorder. *Bioinformatics* **20**:2138-2139.
4. McGuffin,L.J. (2008). Intrinsic disorder prediction from the analysis of multiple protein fold recognition models. *Bioinformatics* **24**:1798-1804.
5. Dosztányi,Z., et al. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21**:3433-3434.

BioSerf

Server-based de novo and fold recognition predictions using BioSerf

S.M. Ward[1], D.W. Buchan[1], and D.T. Jones[1]

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

d.jones@cs.ucl.ac.uk

URL: http://bioinf.cs.ucl.ac.uk

The UCL BioSerf server implements a fully automated template selection and hybrid homology/*de novo* modelling strategy.

Methods

BioSerf initially attempts to find an appropriate template for homology modelling, if that process fails it switches to *de novo* modelling. Template selection uses a range of our algorithms, including PSIPRED [1] and pGenTHREADER [2], to attempt to 'intelligently' select appropriate homology modelling templates with a given homology threshold. On selection of a valid template or templates Bioserf then uses MODELLER [3] to build an appropriate model. Should template selection fail to find a suitable template or if it target sequence fails to achieve sufficient homology coverage with the potential templates then FRAGFOLD [4] is instead used to build a *de novo* model

Availability

BioSerf can be access from the following URL: http://bioinf.cs.ucl.ac.uk/bio_serf/public_job

1. Jones, D.T. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol (1999), 292, 195–202.
2. Lobley, A., Sadowski, M.I. and Jones, D.T. (2009) pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination, Bioinformatics, 25, 1761-1767.
3. Eswar, N., Eramian, D., Webb, B., Shen, M.Y. & Sali, A. Protein structure modeling with MODELLER. Methods Mol Biol (2008), **426**, 145–159.
4. 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.

bujnicki-kolinski

Protein structure prediction by CABS and TRACER with restraints derived from MQAP-scored models.

M. Jamroz[1], J.M. Bujnicki[2,3] K. Mikołajczak[3], P. Wojciechowski[2], A. Koliński[1]

[1]*Laboratory of Theory of Biopolymers, Faculty of Chemistry, University of Warsaw, ul. Pasteura 1, 02-093 Warsaw, Poland*

[2]*Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology, ul. Trojdena 4, 02-109 Warsaw, Poland,*

[3]*Laboratory of Bioinformatics, Institute of Molecular Biology and Biotechnology, Faculty of Biology, ul. Umultowska 89, 61-614, Poznan, Poland,*

To predict the tertiary structure of full-length sequences of all targets in CASP9, we used de novo folding with a lattice-based protein modeling tool CABS[1] developed in the Kolinski group, based on restraints derived from various alternative models.

The GeneSilico metaserver[2] was used to identify domains, predict secondary structure, and generate fold recognition (FR) alignments. These FR alignments were converted to full-atom models using MODELLER[3] and/or SWISS-MODEL[4], frequently involving optimization according to the "FRankenstein's Monster" approach[5]. Additionally, we downloaded all server models from the CASP website.

All these models were evaluated by the newest version of MetaMQAP[6] and other MQAPs developed in the Bujnicki group, which included global and local evaluation for individual models as well as based on clustering (see the abstract by Pawlowski et al.). 3-50 best-scoring models, depending on the overall predicted quality, were used as sources of pairwise restraints, with the strength of restraints depending on the predicted quality of individual residues.

The newest implementation of CABS was used to carry out folding guided by restraints on the pairwise distances and secondary structure with the Replica Exchange Monte Carlo sampling technique. In the case of targets with confidently predicted fold, but very uncertain alignments, we generated additional restraints for target-template correspondencies on the level of individual residues and have run simulations by forcing the target chain to thread the backbone of the selected template, using the TRACER method[7].

Decoys generated in the course of simulation were subject to the average linkage hierarchical clustering. For representative decoys from each cluster, full-atom models were rebuilt, and re-scored with MetaMQAP[6] and DFIRE[8]. Five models were selected for submission based on combination of various criteria, including the MQAP scores, the size, density, and average energy of the corresponding clusters, the visual evaluation of the full-atom structures and their relationship to the original templates.

1. Kolinski A. Protein modeling and structure prediction with a reduced representation. Acta Biochim Pol 2004;51(2):349-371.
2. Kurowski MA, Bujnicki JM. GeneSilico protein structure prediction meta-server. Nucleic Acids Res 2003;31(13):3305-3307.
3. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. Journal of molecular biology 1993;234(3):779-815.
4. Arnold K, Bordoli L, Kopp J, Schwede T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics (Oxford, England) 2006;22(2):195-201.

5. Kosinski J, Cymerman IA, Feder M, Kurowski MA, Sasin JM, Bujnicki JM. 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 2003;53 Suppl 6:369-379.
6. Pawlowski M, Gajda MJ, Matlak R, Bujnicki JM. MetaMQAP: a meta-server for the quality assessment of protein models BMC Bioinformatics 2008;9(1):403.
7. Trojanowski S, Rutkowska A, Kolinski A. TRACER. A new approach to comparative modeling that combines threading with free-space conformational sampling. Acta Biochim Pol 2010;57(1):125-133.
8. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 2002;11(11):2714-2726.
9. Boniecki M, Rotkiewicz P, Skolnick J, Kolinski A. Protein fragment reconstruction using various modeling techniques. Journal of computer-aided molecular design 2003;17(11):725-738.

CBRC_POODLE

POODLE-I: Prediction of disordered region by integrating POODLE series based on workflow approach

S. Hirose, K. Shimizu and T. Noguchi
Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, Japan
poodle@cbrc.jp

POODLE-I ("I" stands for Integration) is a disordered region prediction server that combined the prediction results of POODLE series and structural information predictors by application of a workflow approach[1]. POODLE series consists of three predictors, POODLE-S[2], -L[3], and –W[4], that they target different disordered region according to their length. In POODLE-I, structural information predictors employs PSIPRED[5], jpred[6], and sable[7] as secondary structure prediction, jpred and sable as accessible surface area prediction, genThreader[8], and HHpred[9] as fold recognition, and COILS[10] as coiled coil region prediction.

We assumed that the factor causing a short disordered region might be different from the factor causing long one: a short disordered region is mainly determined according to whether it is located within a structure such as a loop or linker. By contrast, the long disordered region is mainly affected to the physic-chemical property derived from the primary sequence such as low hydrophobicity or high net charge. Based on the idea, the disordered region prediction flow is divisible into two parts. One flow predicts a long disordered region based on physic-chemical property of amino acid, the other predicts short disordered region based on information of protein structure. Initially, POODLE-L and -W were executed. If a long disordered region was predicted in a query, the both termini of it were modified by considering prediction result of secondary structure and coiled coil. Then, POODLE-S was executed for the ordered region predicted by POODLE-L and –W. If a disordered region was predicted it was confirmed using results of structural information predictors. Finally, all residues were labeled as whether an ordered or disordered region.

As an experiment in CASP9, flexible region predictor[11], which predicts the degree of motions derived from Normal Mode Analysis (NMA), was added in the workflow.

Availability

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

1. Hirose,S., Shimizu,K. & Noguchi,T. (2010). POODLE-I: Disordered region prediction by integrating POODLE series and structural information predictors based on a workflow approach. *In Silico Biology* **10**, 0015.
2. Shimizu,K., Hirose,S. & Noguchi,T. (2007). POODLE-S: web application for predicting protein disorder by using physicochemical features and reduced amino acid set of a protein-specific scoring matrix. *Bioinformatics* **23**(17), 2337-2338.
3. Hirose,S., Shimizu,K., Kanai,S., Kuroda,Y. & Noguchi,T. (2007). POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions. *Bioinformatics* **23**(17), 2046-2053.
4. Shimizu,K., Muraoka,Y., Hirose,S., Tomii,K. & Noguchi,T. (2007). Predicting mostly disordered proteins by using structure-unknown protein data. *BMC Bioinformatics* **8**, 78.
5. McGuffin,J., Bryson,K. & Jones,D.T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics* **16**(4), 404-405.

6. Cuff,J.A. & Barton,G.J. (2000). Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins* **40**(3), 502-511.
7. Adamczak,R., Porollo,A. & Meller,J. (2005). Combining prediction of secondary structure and solvent accessibility in proteins. *Proteins* **59**(3), 467-475.
8. Jones D.T. (1999). GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* **287**(4), 797-815.
9. Söding,J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**(7), 951-960.
10. Lupas,A., Van Dyke,M. & Stock,J. (1991). Predicting coiled coils from protein sequence. *Science* **252**(5009), 1162-1164.
11. Hirose,S., Yokota,K., Kuroda,Y., Wako,H., Endo,S., Kanai,S. & Noguchi T. (2010). Prediction of protein motions from amino acid sequence and its application to protein-protein interaction. *BMC Structural Biology* **10**, 20.

Chicken_George

Protein structure prediction with SimFold in CASP9

S. Minami[1], K. Sawada[2], and G. Chikenji[2]
[1] - *Department of Computational Science and Engineering, Graduate School of Engineering, Nagoya University*
[2] – *Department of Applied Physics, Graduate School of Engineering, Nagoya University*
chikenji@tbp.nuap.nagoya-u.ac.jp

Our group, Chicken_George, submitted predictions for all the human/server targets in the tertiary structure category. However, we mainly concentrated our efforts in Medium/Hard targets. Our primary tools used in CASP9 are SimFold energy function for evaluation of structure qualities, fragment assembly method for generating or refining structures, and consensus based method for template identification.

The protein structure representation and energy function

We use SimFold[1] as a primary tool for Medium/Hard targets. It is a protein structure prediction toolbox that we have been developing. SimFold uses a reduced protein structure representation that has explicit backbone atoms and a sphere at the center of mass of side-chain atoms. The energy function consists of several terms such as hydrophobic interaction, hydrogen bonding, and so on. Their functional forms are well based on physico-chemical consideration so that each energetic term can be interpreted as a physical force. The explicit expression of the energy function was previously described[1].

Template identification and target classification

For template identification, we used one of the consensus based methods, ModFOLDclust[2], using all the CASP server predictions. ModFOLDclust was also used as a classfier of target difficulties. Targets are classified as Easy, Medium or Hard, when the ModFOLDclust score of the top scoring model is >0.6, 2.0<score<0.6, or <0.2, respectively. For easy targets, we simply submitted top scoring models without any modification.

Medium targets

Our attempts for medium targets were to refine the template structure identified by ModFOLDclust, and to model unreliable regions by *De Novo* protocols. The template structure we used as a starting point was generally the top scoring model identified by ModFOLDclust from CASP server predictions. Per-residue model quality assessment was also performed in order to identify unreliable regions. We assumed those regions with low local model quality score (< 0.4) as unreliable. For some difficult targets, alignments of a starting point template were manually refined by hand. Then, we performed fragment assembly *De Novo* modeling by SimFold software for unreliable regions with the other regions fixed. Fragment candidates were prepared by profile-profile comparison. We also used fragments extracted from templates collected by the structure alignment program[3] based on the starting point template.

Hard targets.

For hard targets, we didn't use any information of CASP server predictions. Instead, fragment assembly *De Novo* prediction was performed with SimFold energy function. Submitted models were basically selected from low energy structures by structure clustering. If we failed to obtain sufficient cluster size, we selected models by visual inspection. For some very hard targets, we also folded some sequence homologues of the target.

1. Fujitsuka, Y., Chikenji, G. & Takada, S. (2006). SimFold energy function for de novo protein structure prediction: Consensus with Rosetta. *Proteins.* **62**, 381-398.
2. McGuffin, L.J. (2009), *Proteins.* **77**, 185-190.
3. Zhang, Y. & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic. Acids. Res.* **33**, 2302-2309.

chuo-fams

Construction of the Function for Protein Structure Prediction and the Homology Modeling System

Mami Arai[1], Kazuhiko Kanou[2], Genki Terashi[3], Hideaki Umeyama[3], Mitsuo Iwadate[4]
[1] *Physics Course, Graduate School of Science and Engineering, Chuo University; 1–13–27 Kasuga, Bunkyo-ku, Tokyo 112–8551, Japan.*
[2] *Infectious Disease Sureillance Center, National Institute of Infectious Diseases; 1-23-1 Toyama, Shinjuku-ku, Tokyo, 162-8640,Japan.*
[3] *School of Pharmacy, Kitasato University; 5–9–1 Shirokane, Minato-ku, Tokyo 108–8641, Japan.*
[4] *Department of Biological Sciences, Faculty of Science and Engineering, Chuo University.*
n29002@educ.kc.chuo-u.ac.jp

Our server team "chuo-fams" attended TS category in CASP9. For attending CASP9, We constructed scoreA the Function for protein structure prediction and the homology modeling system based on the experience of past CASP. We used the modeling software FAMS[1] and the homology search software PSI-BLAST, HHsearch, SPARKS2, SP3, HMMER and HHM_BLAST. Forpredicting the model similarity from the alignment, we suggest the new scoreing function based on PF_score[2]. We constructed the homology modeling system incorporating the new scoring function and the principal component analysis.

In this report, we report the scoreA and the homology modeling system removing the error of the system in CASP9.

Methods

scoreA is the prediction function of the modeling accuracy, suggested based on PF_score. The arguments of scoreA function are following three parameters, *length* the amino acid sequence length of model, *align_score* the score representing the homology of alignment calculated by BLOSUM62 and affine gap penalty (opening -10, extended -1) and *ss_score* the score representing the rate of secondary structure identity between PSIPRED[3] prediction of target and STRIDE[4] judgment of template protein PDB. For comparing PSIPRED and STRIDE, we executed both programs to the 95% non-redundant 18300 sequences of PDB (May 2, 2008). And we counted the number of residues of each secondary structure combination. The counted numbers were converted to the rates *P(X)*. Therefore we use the odds value calculated by following eq.(1) as the score matrix for *ss_score* (table.1).

$$\log\left(\frac{P(C:T)}{Pp(C)\times Ps(T)}\right)$$

P(C:T) : The appearance ratio of pair of C and T in all pairs.
Pp(C) : The appearance ratio of C in the PSIPRED sequence.
Ps(T) : The appearance ratio of T in the STRIDE sequence.

eq.(1)

PSI＼STR	H	B	E	G	I	T	C
H	0.891	−1.540	−2.448	0.244	0.497	−1.105	−1.397
E	−2.832	0.022	1.269	−1.428	−1.498	−1.150	−0.577
C	−1.376	0.502	−0.673	0.133	−0.175	0.642	0.616

Table. 1 The score matrix between PSIPRED and STRIDE

The scoreA is calculated by following eq.(2).

$$\text{scoreA} = a \times \text{length} + \text{align\_score} + b \times \text{ss\_score} \quad \text{eq.(2)}$$

Above coefficients *a* and *b* were optimized by the maximization of sum of rateGDT_TS (the rate of GDT_TS for the max GDT_TS in all models of a target). We used the 30% non-redundant 6498 sequences (Apr 25, 2008) with less than the 50% amino acid identity alignments. As a result, optimized coefficients were *a*=1.4 and *b*=3.0 (Fig.1).

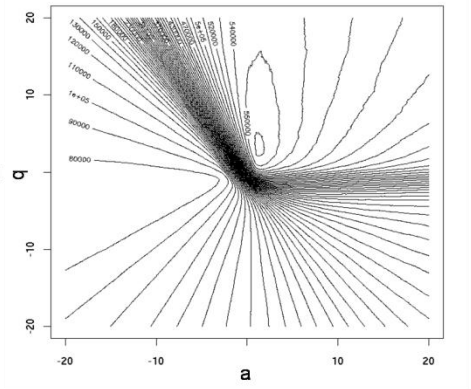


Fig. 1 The contour plot of the sum of rateGDT_TS

We constructed the homology modeling system incorporating the new composite score and the principal component analysis (PCA).

The system executes the following steps.

1. Obtain alignments from homology search software against PDB sequences.
2. Obtain models with FAMS from the alignments.
3. Attempt PCA of alignments and models, plot the result.
4. Attempt the hierarchical clustering of the model plots.
5. Select the representative cluster by an average of CIRCLE[5] (based on physiochemical free energy) value.
6. Select the representative model in the representative cluster by the composite score.
7. Obtain the model for submitting by full-modeling the representative model with FAMS.

The composite score is defined by CIRCLE value and weighted scoreA. The weight is defined 0.07 by maximizing the sum of Z_score of GDT_TS values in CASP8.

In PCA, we determined how many components are used by the threshold 0.83 of the sum of contribution ratios in PCA. The threshold was determined by maximizing the sum of Z_score of GDT_TS values through the system for CASP8 (Fig.2). It is aim that the noise data is removed by the threshold.

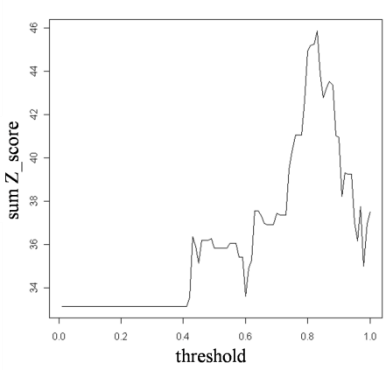


Fig. 2 The relationship between the threshold of the sum of contribution ratios and the sum of Z_score of GDT_TS

Results

For CASP8, we compared the results of scoreA, CIRCLE, the composite score and the constructed system in each difficulty category (TBM-HA, TBM, TBM/FM, FM) (Fig.3).

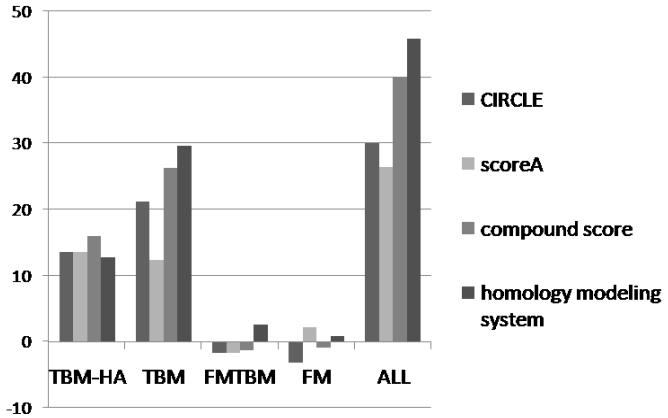


Fig. 3 The bar graph of the sum of Z_score (CASP8)

These scores selected the model of the highest value. In all target (ALL) the constructed system got best result in all scoring procedures, but not all good in each category. Therefore, we needed to divide by target difficulty. In PCA, removing the noise by the threshold had higher effect than using the all data. As a result, the noise data was removed effectively.

In this report, we want to assess scoreA function and the homology modeling system including PCA technique to apply the CASP9 targets excluded from the learning set.

Availability

The service of this work is under construction.

1. Ogata K, 2000 Jun;18(3):258-72, 305-6.
2. Iwadate M., CHEMICAL & PHARMACEUTICAL BULLETIN 58(1), 1-10 (2010).
3. Jones, D.T., J. Mol. Biol. 292:195-202 (1999).
4. Heinig, M., Nucl. Acids Res., 32, W500-2 (2004).
5. Terashi G., Proteins, 69, Suppl 8, 98-107 (2007).

chunk-TASSER

Chunk-TASSER server for protein structure prediction in CASP9

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

The chunk-TASSER server[1] participated CASP experiment for the first time in CASP9. We have made several updates to the original chunk-TASSER and the version used by pro-sp3-TASSER[2] and METATASSER[3].

Method

Chunk-TASSER server uses an updated version of the SP^3 threading method[4]. SP^3 updates include filtering of PSIBLAST hit sequences to less than 90% and 70% sequence identity to each other in profile generation with PSIBLAST e-value cutoffs of 0.001 and 1, respectively. For Easy and Medium targets (SP^3 Z-score > 6.0, 4.5 <= Z-score <= 6.0, respectively), templates are ranked by four different scores from the SP^3 output. The four scores are: (1) raw threading score minus the reverse threading score; (2) raw threading score;(3) raw threading score/alignment length; (4) raw threading score/target length. For Hard targets (SP^3 Z-score < 4.5), in addition to threading template models, we also generated full length ab initio models by fragment assembly[5] if the target size is < 200 residues. Threading models and the ab initio models are ranked by FTCOM[6] and the top 20 models are fed into TASSER[7] for refinement. As in original chunk-TASSER, for Medium/Hard targets, chunk models generated by ab initio method are also included in TASSER refinement. A single TASSER run was performed for each target, and the top five SPICKER cluster centroid-based models were used for prediction. Ideal geometry backbone models are then built from the C_α-only cluster centroid models, followed by relaxation/optimization using the TASSER energy and H-bond count. An in-house template-based side-chain building procedure was employed to build the side-chains of the submitted models.

Results

Chunk-TASSER is developed mainly for improving prediction accuracy of Medium/Hard targets. Nevertheless, it also has comparable performance to the top performing servers for Easy targets. Chunk-TASSER server models have good geometry and H-bond score comparable to those of other top performing servers. It is among the top predictors for human/Hard targets, according to unofficial assessment at http://zhanglab.ccmb.med.umich.edu/casp9/.

Availability

The chunk-TASSER program and web service are available at http://cssb.biology.gatech.edu/

1. Zhou, H and Skolnick, J. (2007) Ab initio protein structure prediction using chunk-TASSER.. Biophysical Journal. **9**3,1510-8.
2. Zhou,H and Skolnick, J. (2009) Protein structure prediction by pro-sp3-TASSER. Biophysical Journal. **96**, 2119-27.
3. H. Zhou, S. B. Pandit and J. Skolnick (2009) Performance of the Pro-sp3-TASSER Server in CASP8.

Proteins **77(S9**), 123-127.
4. Zhou, H. and Zhou,H. (2005) Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins **58**, 321--328.
5. Simons,K. et al (2000) Assembly of protein tertiary structure from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J.Mol. Biol. **268**,209-225.
6. Zhou,H and Skolnick, J. (2010) Improving threading algorithms for remote homology modeling by combining fragment and template comparisons. Proteins. **78**, 2041-8.
7. Zhang, Y. and J. Skolnick (2004) Automated structure prediction of weakly homologous proteins on genomic scale. Proc. Natl. Acad. Sci. (USA) **101**,7594--7599.

Circle

Template based modeling server with Model Quality Assessment Program circle

Genki Terashi[1], Kazuhiko Kanou[1,2], Makoto Oosawa[1], Yuuki Nakamura[1], Hideaki Umeyama[1], and Mayuko Takeda-Shitaka[1]
[1] - School of Pharmacy, Kitasato University
2- Infectious Disease Surveillance Center, National Institute of Infectious Disease
terashig@pharm.kitasato-u.ac.jp

In 9th round of CASP, we have developed template based modeling server circle with CIRCLE QA program[1]. The circle aims at identifying the near native models and incorrect model from 50~60 generated candidates without using clustering based methods (no-consensus).

Methods

1. Alignments search and Template based modeling

A target sequence was searched against PDB sequence database by BLAST, PSI-BLAST, PSF-BLAST, RPS-BLAST, IMPALA, Pfam-BLAST, CSI-BLAST, SPARKS2, SP3 and HHsearch. Various alignments were filtered with its alignments score according to our FAMSD[2] procedure. Then, the 50~60 alignments were fed into template based modeling program FAMS[3] to generate 50~60 full atom models.

2. Model evaluation by CIRCLE QA program

The all candidate models were evaluated and ranked by CIRCLE QA program. CIRCLE considers two terms for the model quality as:

$$TotalScore = \begin{cases} \sum_{n}^{length} \left(0.35 \cdot SSscore + SideChainScore_{EASY}\right)_n & EASY \\ \sum_{n}^{length} \left(0.75 \cdot SSscore + SideChainScore_{HARD}\right)_n & HARD \end{cases} \quad (1)$$

Where *SideChainScore* represents the quality score of side-chain coordinates calculated from the side-chain environment of each residue and *SSscore* is the similarity between the secondary structure propensities predicted for an amino acid sequence by PSI-PRED and the secondary structure of the three-dimensional model. The side-chain environment for each residue is determined from the fraction of the surface area of the side-chain covered by the polar atoms, the fraction of the side-chain area buried by any other atoms, and the secondary structure.

1. Terashi G, Takeda-Shitaka M, Kanou K, Iwadate M, Takaya D, Hosoi A, Ohta K, and Umeyama H (2007). Fams-ace: a combined method to select the best model after remodeling all server models. *Proteins*.69 Suppl 8:98-107.
2. Kanou K, Iwadate M, Hirata T, Terashi G, Umeyama H, Takeda-Shitaka M. Chem Pharm Bull (Tokyo). 2009 Dec;57(12):1335-42.
3. Ogata, K. and Umeyama, H. (2000) An automatic homology modeling method consisting of database searches and simulated annealing. *J. Mol. Graphics Mod*. 18(3):258-72, 305-6.

CNIO

Using predicted contacts to select model structures

I. Ezkurdia[1] and M.L. Tress[1]
CNIO *(Spanish National Cancer Research Centre), Madrid, Spain*
mtress@cnio.es

During the 7th Critical Assessment of Protein Structure Prediction (CASP7) experiment, it was suggested that the real value of predicted residue–residue contacts might lie in the scoring of 3D model structures (1). In a follow-up work (2) we showed that the information contained in the predicted residue–residue contacts would probably help in the selection of 3D models in the free modelling regime and for the harder comparative modelling targets. Indeed, we found that in many cases, the models selected using just predicted contacts had better GDT-TS scores than all but the best 3D prediction groups. This selective power of predicted contacts was surprising because of the well-known low accuracy of residue–residue contact predictions.

Here, we wanted to put predicted contacts to the test in a blind experiment. Are contact prediction methods able to aid in the selection of 3D structural models? We used three available contact prediction methods and attempted to use them to score models predicted by the structure prediction servers.

Methods

Initially we used seven methods to predict contacts. However the three methods from the SAM server (3) were not able to return results in time for most targets. The remaining four methods were installed locally. Three of the remaining four methods were sequence-based; svmseq (4) and svmcon (5) predicted for all targets, but nncon (6) generally only provided predictions when templates could be found. The final method generated consensus contacts from the models predicted by the CASP9 servers for each target.

For each method we set a optimal sequence separation cut-off and constant number of predicted contacts to use in the predictions. We used these predicted contacts to chose the best scoring server models for each target. The winning server model for each contact predictor was simply the model with the least distance between all the predicted contacts as in the paper (2). No other method apart from predicted contacts was used to select models, however we did throw out those models with too many clashes. We also found that there were many models with long and compacted beta-strands and these models had an advantage when assessed with predicted contacts. Several methods appeared to predict long and compacted beta-strands quite frequently, so we inspected the chosen models by eye in order to remove these models too.

We made predictions only for the harder targets. We defined the harder targets as those targets for which we could not detect templates with PSI-BLAST (7).

Results

See CASP assessors.

1. Izarzugaza J, Grana O, Tress ML, Valencia A and Clarke N (2007). Assessment of intramolecular contact predictions for CASP7. *Proteins*, **69**, 152-158.

2. Tress ML and Valencia A (2010). Predicted residue-residue contacts can help the scoring of 3D models. *Proteins*, **78**, 1980-1991
3. Shackelford G and Karplus K (2007). Contact prediction using mutual information and neural nets. *Proteins*, **69**, 159-164.
4. Wu S and Zhang Y (2008). A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. *Bioinformatics*, **24**, 924-931.
5. Cheng J and Baldi P (2007). Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinformatics*, **8**, 113.
6. Tegge AN, Wang Z, Eickholt J and Cheng J (2009). NNcon: Improved protein contact map prediction using 2D-recursive neural networks. *Nucleic Acids Res*, **37**, W515-W518.
7. Altschul S, Madden T, Schäffer A, Zhang J, Zhang Z, Miller W and Lipman D (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389-3402.

ConFuzz

ConFuzz residue-residue proximity prediction metaserver

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

ConFuzz predictor is a weighed filtered consensus approach to contact map (RR) prediction. In order to generate weights to external methods, Sw-score[1], adapted to handle contacts, was calculated per each method against protein benchmark set. Furthermore, additional protocol was used to filter residua that tend to generate too much noise (i.e. expected number of contacts which fairly exceeds physical or statistical limits). After exclusion of erroneous residua in each method (and all their contacts), consensus prediction was built and final contact map was generated.

Methods

In order to seed the consensus approach, four external methods were employed: NNcon[2], SVMcon[3], SVMseq[4] and PoCM[5]. The main server routine was written in the Python programming language (Python2.6). The server used for predictions allowed up to eight concurrent prediction threads. In most cases, all four predictions were processed in parallel which resulted in relatively fast calculation. After all the methods finish their prediction routine, the server reads out all contact maps generated in various formats and passes them to the filter routine.

Filtering redundant contacts (and residua) is realized by summing up all contact per each residue (the so-called contact number). The predicted contact maps are probabilistic (fuzzy), though one might calculate the expected contact number as well. In our previous attempts it turned out that for many residua predictions exceeded statistical or physical bounds for contact number, reaching figures such as 30 contacts per residue (or more). We conducted manual inspections of predicted maps using PROTMAP2D[6] software for visualizing fuzzy contact map. Then we decided to set the threshold to 20, which is more permissive than strict statistical values, which in e.g. case of C-beta maps should not exceed 18 contacts per residue. Thus, for residues with more than 20 predicted contacts, all contacts were deleted .

The consensus prediction was calculated by simply averaging values of contact probability, weighed by SW-score as defined for CASP6 disorder prediction assessment. The score was precomputed for each predictor using a test set of 362 proteins generated by the PISCES[7] server. In order to assess statistical measures like TP (true positive) or TN (true negative), we assumed according to the literature, 10*L as expected number of contacts where L is the length of the protein. Thus all predicted contacts were sorted by their probability and divided into top 10*L as predicted contacts while the rest were considered predicted as non-contacts, regardless of the prediction value.

In the end, all the predictions were scaled to ensure all values for residue pairs predicted “in contact” are above 0.5. Additionally, values were rounded and rescaled, to follow CASP standards. Note that most of the external predictors used predict C-alpha contacts, despite C-beta being CASP RR standard. But since no sound protocol is known to convert those contacts, we used the notion that each C-alpha contact is neighboring its respective C-beta contact in 2D map. With this assumption, altogether with the fact that practically all predicted maps were filled with nearly all positive values, we decided that

the eventual loss of this approach is only lowering/lifting the actual probability value when switching from C-alpha to C-beta image.

1. Jin,Y. & Dunbrack Jr,R.L. (2005). Assessment of disorder predictions in CASP6. *Proteins: Structure, Function, Bioinformatics* **61**(S7):167-175.
2. Tegge,A.N., Wang,Z., Eickholt,J. & Cheng,J. (2009). NNcon: improved protein contact map prediction using 2D-recursive neural networks. *Nucleic Acids Res.* **37**(2):515-518.
3. Cheng,J. & Baldi,P. (2007). Improved contact prediction using support vector machines and a large feature set. *BMC Bioinformatics* **8**:113.
4. Wu,S. & Zhang,Y. (2008). A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. *Bioinformatics* (Oxford, England) **24**(7):924-931.
5. Hamilton,N., Burrage,K., Ragan,M.A. & Huber,T. (2004). Protein contact prediction using patterns of correlation. *Proteins: Structure, Function, Bioinformatics* **56**(4):679-684.
6. Pietal,M.J., Tuszynska,I. & Bujnicki,J.M. (2007). PROTMAP2D: visualization, comparison and analysis of 2D maps of protein structure. *Bioinformatics* (Oxford, England) **23**(11):1429-1430.
7. Wang,G. & Dunbrack Jr,R.L. (2003). PISCES: a protein sequence culling server. *Bioinformatics* (Oxford, England) **19**(12):1589-1591.

ConQuass

ConQuass: using evolutionary conservation for quality assessment of protein model structures

M. Kalman and N. Ben-Tal
Department of Biochemistry and Molecular Biology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel
nirb@tauex.tau.ac.il

Programs that evaluate the quality of a protein structural model are important both for validating the structure determination procedure and for guiding the model-building process. Such programs are based on properties of native structures that are generally not expected for faulty models. One such property, which is rarely used for automatic structure quality assessment, is the tendency for evolutionarily conserved residues to be located at the structural core and for variable residues to be located at the surface. We have developed a new very simple Model Quality Assessment Program (MQAP) called ConQuass[1] (Conservation-based Quality Assessment), which is based solely on the correlation between each residue's degree of evolutionary conservation and its accessibility in the structure

Methods

The conservation pattern for each target was calculated using ConSurf[2], with 3 PSI-BLAST[3] iterations, maximum of 300 sequences, and the alignment performed using MUSCLE[4]. The solvent accessibility of each residue in each model was calculated using NACCESS[5]. The ConQuass score quantifies the compatibility of the accessibility and conservation patterns, and was calculated using the same protocol as in the original publication[1].

Results

We checked the performance of ConQuass on the CASP8 dataset[6]. Surprisingly, when reliable evolutionary conservation information exists, this single feature can achieve performance that is quite comparable to other pure single-structure quality assessment programs, such as QMEAN[7] and MULTICOM[8], that are based on the integration of many different structural features. We also showed that ConQuass is complementary to these existing methods and could potentially be integrated with them to improve the overall performance. This highlights the importance of this feature both for use in quality assessment programs, and for integration in structure prediction schemes. As an MQAP, ConQuass offers the advantage of giving easily interpretable results, as the score is based on a single straightforward feature.

We also performed a preliminary analysis of ConQuass' performance on the CASP9 targets whose native structures were already published. We found the mean per-target Pearson correlation between the ConQuass score and the GDT-TS to be 0.644. The correlation was higher (0.812) when considering only the targets with reliable conservation information. It is important to note that these targets were selected in advance (i.e., before the experimental structure was known), and this selection was noted in the REMARK record of the submitted predictions. Overall, the performance of ConQuass in CASP9 was comparable to its performance on the CASP8 and CASP7 datasets[1].

Availability

ConQuass is freely available for academic use, both as a downloadable program and as a web server, at http://bental.tau.ac.il/ConQuass/.

1. Kalman,M. & Ben-Tal,N. (2010). Quality assessment of protein model-structures using evolutionary conservation. *Bioinformatics* **26**, 1299-1307.
2. Ashkenazy,H., Erez,E., Martz,E., Pupko,T. & Ben-Tal,N. (2010) ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res.* **38 Supply**, W529-533.
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. Edgar,R.C. (2004) MUSCLE: multiple sequence alignments with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792-1797.
5. Hubbard,S.J. & Thornton,J.M. (1993) 'NACCESS', computer program.
6. Moult,J., Fidelis,K., Kryshtafovych,A., Rost,B. & Tramontano,A. (2009) Critical assessment of methods of protein structure prediction – Round VIII. *Proteins* **77 Suppl 9**, 1-4.
7. Benkert,P., Tosatto,S.C. & Schwede,T. (2009) Global and local model quality estimation at CASP8 using the scoring functions QMEAN and QMEANclust. *Proteins* **77 Suppl 9**, 173-180.
8. Cheng,J., Wang,Z., Tegge,A.N. & Eickholt,J. (2009) Prediction of global and local quality of CASP8 models by MULTICOM series. *Proteins* **77 Suppl 9**, 181-184.

cpu_hsfang

Identification of native-like protein structures among sets of decoys employing a novel average measures approach

Juan Li[1▲], Huisheng Fang[2▲*], Cheng Guo[2], Guanhua Ai[2], Hongqian Yang[2], Jianhong Zhou[2], Nanxi Huang[2], Xiangzhen Li[2], Xiaoyan Xu[2], Xingbao Qi[2], Ying Zhang[2], Zipeng Liu[2], Kaixian Chen[3★]

[1] - Department of Blood, Nanjing Drum Tower Hospital, Zhongshan Road 321 Nanjing 210008, [2] - School of Life Science and Technology, China Pharmaceutical University Tongjia Xiang 24, Nanjing 210009, China, [3] - Shanghai University of Traditional Chinese Medicine , 1200 Cailun Road, Zhangjiang Hi-Tech Park, Pudong New District, Shanghai, 201203, China

*Corresponding Author Email Address: hsfang889@163.com

▲Juan Li and Huisheng Fang contributed equally to this work

★General Designer

Most protein structure predictors involve initial generation of a large collection of possible conformations (decoys), among which are native or near-native conformations[1]. Therefore, accurate identification of native-like conformations among the decoys generated is essential. In fact, recent investigations on CASP have revealed that in most cases, the so-called "best" models selected were, in fact, not the best due to the limitation of the approaches employed to identify native-like conformations among the decoy sets[2]. In the present investigation, we have developed a simple and effective procedure for the recognition of near-native conformations in sets of decoys, employing average rmsd (armsd), the most common measure of similarity between models, and the average alignment score (AAS), an additional common measure developed by Levitt and Gerstein[3].

Methods

In this study, three different kinds of decoy sets were used, i.e. "4state_reduced (Park-Levitt) sets", "fisa (Simons) sets" (http://dd.stanford.edu/) [4] and decoy sets generated via Rosetta (referred to here as Baker sets). We obtained a total of 104 sets of protein decoys possessing the basic features of ab initio sets. Then, our group employed two different average measures (i.e. the average rmsd (armsd) and average alignment score (AAS)) to identify native-like protein structures between the model in question and the other models within the same set of decoys.

Results

The three different types of decoy sets mentioned above were used to evaluate this approach. Comparison of model quality to average measure revealed a significant correlation between these parameters, that average measure can be used effectively for identification of native-like protein models. Moreover, the performance of both armsd and AAS in our experiment was more reliable than that of clustering. The performance of armsd was slightly better than that of AAS. In light of the fact that there are numerous other measures for assessing the similarity between protein structures, other analogous approaches to the identification of native-like protein structures will probably also prove to be useful. Finally, the results of predicting the targets in CASP6 and CASP7 showed that its performance was better than the other servers in these two CASPs

Availability

The server is now in preparation.

1. Li,J., Fang,H. & Chen,K. A transformation of substitution matrix for improving local sequence alignment quality for distantly related protein. (In Preparation)
2. Skolnick,J., Kolinski,A., Kihara,D., Betancourt,M., Rotkiewicz,P. & Boniecki,M. (2001) Ab initio protein structure prediction via a combination of threading, lattice folding, clustering, and structure refinement. *Proteins*. **Suppl 5**, 145-156.
3. Levitt,M. & Gerstein,M. (1998) Comprehensive assessment of automatic structural alignment against a manual standard, the scop classification of proteins. *Protein Sci*. **7**, 445-456.
4. Shortle,D., Simons,K.T. & Baker,D. (1998) Clustering of low-energy conformations near the native structures of small proteins. *Proc. Natl. Acad. Sci, USA.* **95**, 11158-11162.

DCLAB

Combining Spectral Analysis with Motif Search and Homology Modeling for Protein Structure Prediction

Carlos A. Del Carpio[1], Daishi Kitazawa[1], Masatake Sugita,[2] Eiichiro Ichiishi[3] and Masa Toshi Yamamoto[1]

[1] *Drosophila Genetic Resources Center. Kyoto Institute of Technology. Japan.*

[2] *Dept Bioinformatics, Col Bioinformatics, Ritsmeikan University. Kyoto-Japan* [3] *Japan Advanced Institute of Science and Technology, Kanazawa, Japan.*

carlos@dgrc.kit.ac.jp

To predict protein 3D structures in CASP9 we have combined our original system for protein 3D structure prediction PIPS1,2 with conventional sequence alignment techniques as well as a new methodology for motif search. The underlying concept in the spectral analysis method embedded in PIPS is a periodicity analysis of the physico-chemical properties of the residues constituting a protein primary structure. The analysis is performed using a front-end processing technique in automatic speech recognition[1,2] by means of which the cepstrum (measure of the periodic wiggliness of a frequency response) is computed so as to infer the spectral envelope that depicts the subtle periodicity in physicochemical characteristics of the amino acid sequences. The system extracts a diverse set of proteins from PDB when the methodology is applied to a target sequence in order to search similar folding patterns. Extracted structures rank from scant similarity in terms of amino acid composition to high similarity ones. Then a more specific sequence alignment like FASTA (http://www.ebi.ac.uk/Tools/fasta33) or BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi) can be applied to the reduced set of structures obtained by our spectral oriented methodology. This combined method has shown a fair degree of effectiveness to select optimal templates for a determined target, both in terms of processing times as well as quality of template. The threading algorithm is then pursued by an energy minimization process for the newly built structure. Table 1 shows a list of the targets in which the methodology has succeeded in recognized the closest folding pattern for the targets in CASP9.

	Target	Length	Fit Length	RMSD
1	T0526_3	290	229	1.98
2	T0531_5	65	34	2.23
3	T0566_1	156	126	2.41
4	T0574_1	126	51	2.53
5	T0576_3	172	164	1.96
6	T0580_5	105	68	2.13
7	T0584_3	352	240	2.16
8	T0586_1	125	106	1.9
9	T0588_1	400	239	2.47
10	T0590_3	137	65	2.13
11	T0592_1	144	101	2.03
12	T0594_1	140	130	1.59
13	T0596_1	213	106	2.43
14	T0602_5	123	54	2.25
15	T0605_1	72	45	1.91
16	T0616_1	103	60	2.4
17	T0618_1	182	60	2.4
18	T0622_2	138	56	2.18

Table 1. Comparison of results for some CASP9 targets(Lenght in number of amino acids)

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

Dill

Physics-Based Structure Prediction by Zipping and Assembly

J.L. MacCallum[1], A. Pérez[1], G.C. Rollins[1], J. Lee[2], and K.A. Dill[1]
[1] - *Department of Pharmaceutical Chemistry, University of California San Francisco,*
[2] – *Department of Bioinformatics and Life Science, Soongsil University, Seoul 156-743, Korea*
dill@maxwell.ucsf.edu

Our major aim is to harness the physical principles of protein folding to create faster search strategies for conformational sampling and structure prediction. Our CASP9 pipeline, which we call ZAMDP (Zipping and Assembly Mechanism by Dynamic Programming), combines rigid body assembly moves with molecular dynamics to rapidly navigate conformational space.

In ZAMDP, we assemble secondary structure fragments (predicted by Psipred and Porter) using a rigid body assembly algorithm called MASH. We use MASH in conjunction with CKY, an algorithm adapted from computational linguistics, to investigate all possible assembly pathways. Structures are scored according to Amber96/GBSA and then clustered by RMSD. The best scoring members of each cluster are then refined by replica exchange molecular dynamics or iterative simulated annealing.

MASH (Magical Assembly of Sheets and Helices) is an algorithm for assembling pairs of protein fragments joined by loops[1]. First, we generate a dot surface around each of the two fragments. Then, we select a pair of dots, one on each fragment. We align the protein fragments to put the pair of dots in contact with one another. Finally, we use a robotics-based loop closure algorithm to connect the two fragments[2]. We repeat this procedure to generate an ensemble of possible fragment pairings.

ZAM is a ”local first, global later” view of folding. This view of folding is similar to the way that CKY, an algorithm from computational linguistics, parses sentences[3,4]. We use a data structure, similar to the one that CKY uses to parse sentences, to enumerate and organize the possible fragment assembly pathways. The parse-chart directs the assembly process and breaks it down into a series of pairwise assembly steps, which are carried out by MASH, as described above.

1. Wu,G.A., Coutsias,E.A. & Dill,K.A. (2008) Iterative assembly of helical proteins by optimal hydrophobic packing. *Structure **16***, 1257-1266.
2. Coutsias,E.A., Seok,C., Jacobson,M.P. & Dill,K.A. (2004) A kinematic view of loop closure. *J. Comput. Chem.* **25**, 510-528.
3. Dill,K.A., Lucas,A., Hockenmaier,J., Huang,L., Chiang,D. & Joshi,A.K. (2007) Computational linguistics: A new tool for exploring biopolymer structures and statistical mechanics. *Polymer **48***, 4289-4300.
4. Hockenmaier,J., Joshi,A.K. & Dill,K.A. (2007) Routes are trees: The parsing perspective on protein folding. *Proteins **66***, 1-15.

Distill
Distill_human

Distill: protein structure prediction by Machine Learning

C. Mirabello[1], G. Tradigo[1,2] and G. Pollastri[2]
[1] – UCD Dublin, Ireland, [2] – Università della Magna Grœcia, Italy
gianluca.pollastri@ucd.ie

Distill has two main components: a set of predictors of protein features 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. For CASP9 we have retrained and updated our prediction methods and fold recognition module, and our optimisation algorithm for the conformational search, which now uses "snippets" of PDB structures suggested by our fold recognition algorithm.

The only difference between Distill and Distill_human is that for the latter we evaluated and partially re-ranked Distill's models visually.

Methods

Distill runs 3 rounds of PSI-BLAST against a 90% redundancy reduced UniProt to generate multiple sequence alignments (MSA). The PSSM from the second round is reloaded to search the PDB for templates (e=1e-3). MSA and templates are fed to our 1D prediction systems (all based on BRNN): Porter[1] (secondary structure), PaleAle[4] (solvent accessibility), BrownAle[4] (contact density), Porter+[2] (structural motifs). All predictors use template information as an input alongside the sequence and MSA.

1D predictions are combined into a structural fingerprint[4] (SAMD) which, alongside the PSSM, is used to find remote homologues in the PDB (1-against-all alignment). If this search returns templates that are deemed to be more reliable than the PSI-BLAST ones, all 1D predictions are run again with the new templates as inputs.

In the following stage residue distance and contact maps are predicted by a system based on 2D-Recursive Neural Networks (XXstout[5]). Two types of maps are predicted: binary maps with a contact threshold of 8Å between Cβ, which are submitted to the RR category; 4-class distance maps (thresholds of 8, 13 and 19Å) between Cα which are used for 3D prediction. Inputs for map prediction are: the sequence; MSA; PSI-BLAST and SAMD templates. That is, the maps are template-based whenever suitable templates are found.

The 3D reconstruction, which is only conducted on Cα traces, is run as follows: we run a SAMD search for templates with an e-value of 10,000; for each (overlapping) 9-mer of the protein we gather the structures of the top 50 templates which fully cover it (SAMD_list); a simulated annealing search of the conformational space is run using crankshaft moves to quickly find a minimum of a potential function which rewards formation of predicted contacts; from the previous enpoint a simulated annealing search is run by substituting 9-mers from the conformation with 9-mers from the SAMD_list, and using the same potential function as above.

We run 30 reconstructions for each protein, which we rank by their weighed TM-scores against templates. For the 5 top-ranked models we reconstruct the backbone with Maxsprout (with SABBAC for Distill_human), and the full atoms with Scwrl4. These are the models submitted to CASP.

It should be noted that everything in our pipeline (except BLAST and the software to blow Cα traces into full-atom models) is in house, and that in normal conditions we can provide predictions for a protein in tens of minutes.

Results

For many proteins, our results seem to be competitive to us based on the first 80 structures released, but we await the CASP assessment for this.

Availability

http://dbstill.ucd.ie/distill/

1. Pollastri,G. & McLysaght,A. (2005) Porter, A new, accurate server for protein secondary structure prediction, *Bioinformatics*, **21**(8), 1719–1720.
2. 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.
3. Walsh,I., Martin, A.J.M., Mooney, C., Rubagotti, E., Vullo, A. & Pollastri, G. (2009). Ab initio and homology based prediction of protein domains by recursive neural networks" *BMC Bioinformatics*, **10**,195.
4. Mooney, C. & Pollastri, G. (2009). Beyond the Twilight Zone: Automated prediction of structural properties of proteins by recursive neural networks and remote homology information, *Proteins*, **77**(1), 181-90.
5. Walsh, I., Baú, D., Martin, A.J.M., Mooney, C., Vullo, A. & Pollastri, G. (2009). Ab initio and template-based prediction of multi-class distance maps by two-dimensional recursive neural networks, *BMC Structural Biology*, **9**,5.

Distill_NNPIF

Alberto Jesus Martin

AmMBio group at University College Dublin (Ireland) and
BioComputing UP group at University of Padova (Italy)

Distill_NNPIF is a knowledge-based Model Quality Assessment Program (MQAP) at the residue level which evaluates single protein structure models. Distill_NNPIF also predicts local quality, but it is derived from global quality.

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

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

Dokhlab

Protein Structure Prediction by Ab Initio Folding using Discrete Molecular Dynamics

S. Yin[1], F. Ding[1], J. Hahn[2], P. Kota[1], E. Proctor[1], S. Ramachandran[1], D. Shirvanyants[1], and N.V. Dokholyan[1,3]
[1]*Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill*
[2]*Mathematics Department, Massachusetts Institute of Technology*
[3] dokh@med.unc.edu

Protein structure prediction is difficult when homologous structures are not available for the whole length protein or for the insertion/deletion region. To address this issue, we test an approach involving the *ab initio* folding of either the entire length of the protein for those sequences with no close homolog, or the insertion/deletion fragment regions only in those sequences that do have a close homolog, using all-atom discrete molecular dynamics (DMD) simulation.

Methods

We perform a BLAST[1] search against the PDB to determine if homologous structures exist for a given sequence. If a homologous structure is unavailable, we fold the entire length of the protein from its fully extended conformation using DMD simulation[2]. If homologous structures exist, we use the Medusa[3,4] suite to model point mutations from the template structure, and apply DMD simulation to fold the insertion and deletion segments.

DMD simulations are performed using replica exchange protocols to improve the sampling efficiency. In order to accelerate the *ab inito* folding of the full-length protein, we apply constraints derived from secondary predictions using Psipred[5] and Jpred[6]. We cluster the structures generated by DMD simulations in order to identify a diverse set of candidate conformations. These candidate structures are further optimized using Medusa and ranked based on their final Medusa energy as well as experts' structural quality assessment.

Availability

The implementation of the methods is available for academic users upon request.

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. Ding, F., Tsao, D., Nie, H. & Dokholyan, N.V. (2008). Ab Initio Folding of Proteins with All-Atom Discrete Molecular Dynamics, *Structure*, 16, 1010-1018.
3. Ding, F. & Dokholyan, N.V. (2006). Emergence of Protein Fold Families through Rational Design, *PLoS Comput Biol*, 2, e85-e85.
4. Yin, S.Y., Ding, F. & Dokholyan, N.V. (2007). Eris: an automated estimator of protein stability, *Nature Methods*, 4, 466-467.
5. Jones,D.T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* ***292***, 195-202.
6. Cole,C., Barber, J.D. & Barton, G.J.(2008). The Jpred 3 secondary structure prediction server. *Nucleic Acids Research*, 36, W197-W201.

Elofsson
Pcons
Pcomb
ProQ

Pcons, PconsD, PconsR, PconsM, Pcomb, ProQ and ProQ2 in CASP9

Per Larsson, Arjun Ray, Marcin Skwark, Patrik Björkholm, Erik Lindahl, Björn Wallner and Arne Elofsson
CBR, SBC, DBB, SERC Stockholm University 106 91 Stockholm Sweden
arne@bioinfo.se

Our goal in CASP9 was to extend the use of consensus- (or meta-) server methods, beyond the simple ranking of methods. We have tried three different approaches. In PconsD we have used distance constrains obtained from the models and build new models, while in PconsM we have used traditional multiple alignments as an input to Modeller. In addition a number of refinement protocols were tested in the PconsR method. The performance of these methods were tested against the performance of consensus based MQAPs (Pcons), single model based MQAPs (ProQ and ProQ2) and a combined method (Pcomb).

Methods

In CASP9 a number of prediction methods where assesses. Below follows a short summary of these. Most of the methods were used both for 3D predictions and as MQAPs. The basis for all these predictions is a set of predictions collected by the Pcons.net server[1]. In CASP9 Pcons.net collected predictions from local (blast, rpsblast, HHpred) and external (ffas, forte, hhpred, nfold phyre, Sam-t02) sources. From all these predictions the alignments were obtained and 3D-models were built using Modeller. If no significant templates were found models were also constructed from fragments using Rosetta. It should be noted that all servers did not return predictions in time for all targets. The collected models were then used for the basis for predictions of 3D structures, either just by ranking them or by using the alignments as an input to various schemes as described in the methods section. Further the different scoring functions were also used as MQAPs.

Pcons[2,3] is an MQAP that uses the similarity between a model and all other models for scoring. In CASP9 the scoring is done using the average S-score[4] to all other models.

PconsM[5] is an extension to Pcons where multiple templates are used. Here, the starting point is the highest ranked Pcons model and then additional alignments are added and models (re)-built using Modeller. The models are ranked using ProQ2 (see below). When used as an MQAP the Pcons scoring methodology is used but only the first ranked models are included in the comparison.

PconsD is a novel approach to consensus-based protein structure prediction. Given an ensemble of structure prediction models, it derives a set of structural constraints for a given protein sequence. These in turn are used for building novel protein models. Models in the initial ensemble are scored by a linear combination of Pcons and the new single-model, stand-alone MQAP ProQ2 as in Pcomb. The PconsD MQAP uses the same linear combination as scoring function.

PconsR is a refinement protocol, where the highest ranked Pcons models is refined using 5 different refinement protocols based on MD or MC simulations and different force fields.

ProQ[6] is a non-consensus based MQAP. A number of properties are calculated from a model and then an artificial neural network is used to predict the quality of the model.

ProQ2[7] is an updated improved non-consensus based MQAP. It uses the same properties as ProQ but also includes new features that makes it perform significantly better than ProQ.

Pcomb ranks the different models using a linear combination of ProQ2 and Pcons both when it used to select the top-ranked model and as an MQAP.

Elofsson is our manual predictor. Here Pcomb is applied to the complete set of server predictions.

Further, in the refinement category the given starting structures were refined using Rosetta with particular focus on rebuilding regions that were either said to be poorly refined or that were predicted to be poorly refined by Pcons. At least 10,000 models were constructed for each target and then ProQ, ProQ2, Pcons and Pcomb were applied to rank the models, the lowest Rosetta energy models were also submitted as PconsM.

Results

After the release of 85 targets the sums of GDT_TS for the different methods we evaluated the performance of all our predictions on full-length targets and using the sum of GDT_TS as the evaluations method, see table I. Alternative evaluation protocols gave similar results. First it can be seen that all methods except ProQ performs significantly (up to 7%) better than the best method used as an input to Pcons.net (HHpred). It should however be noted that the models submitted to CASP9 by HHpred directly clearly is better than the models obtained from the HHpred server by Pcons.net. Further two out of the three methods we devised to "go beyond MQAPs" performed slightly better than Pcons. The best performance was obtained by PconsM in CASP9 and the relative improvement over the single template model increase with an increasing number of alignments used, see Figure 1. In addition the novel PconsD method also performed slightly better than Pcons, while the Pcomb method performed on par with Pcons. The manual predictions (Elofsson) performed about 10% better than the methods based on prediction from Pcons.net, however the improvement over the best server here was marginal.

Figure 1: Improvement over model built by a single template in PconsM (line) and the fraction of models submitted using a certain number of alignments (bars).

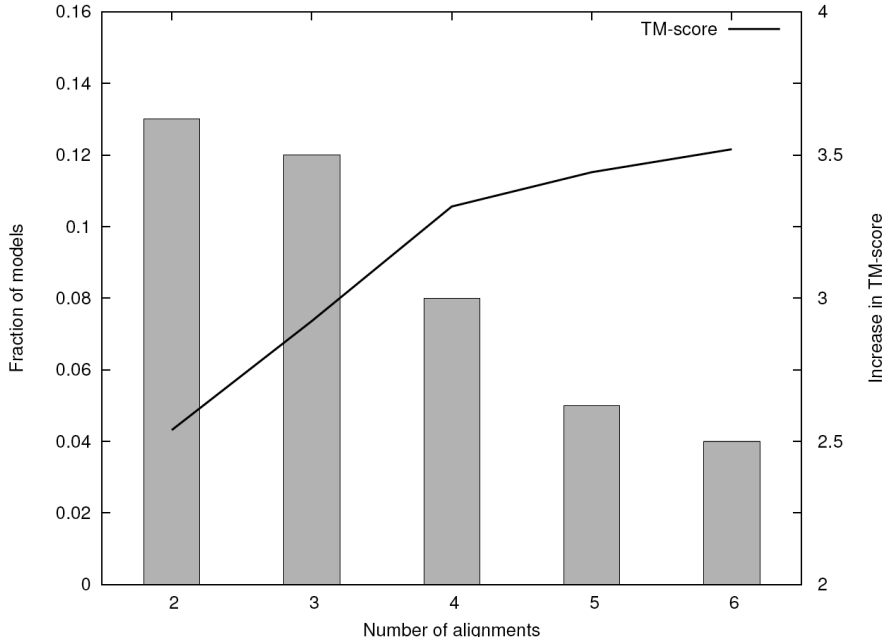


Table I: Performance on 78 full-length targets release in Sep 2010

Method	Sum of GDT_TS	Sum of top 5 GDT_TS
Elofsson	48.6	50.6
PconsM	44.7	46.2
PconsD	44.3	45.7
Pcons	44.0	46.6
Pcomb	44.1	47.2
PconsR	42.0	46.4
ProQ2	43.5	46.6
ProQ	39.0	44.9
HHpred in Pcons.net	41.9	44.1

In addition we tested a number of MQAPs in CASP9. As in earlier years the consensus based MQAPs performed better than the non-consensus based predictors, see Table II. However the most important result was that ProQ2 clearly performs better than the earlier non-consensus based MQAP ProQ.

Table II: MQAP spearman correlations versus TM-score (global) and S-score (local) on 85 full-length targets release before Sep 2010 are reported.

Method	Global	Local
PconsM	0.93	0.83
Pcons	0.94	0.83
Pcomb	0.93	0.83
ProQ2	0.76	0.68
ProQ	0.67	0.50

Availability

The successful methods will become available through the Pcons.net web-interface.

1. Wallner, B., Larsson, P. and Elofsson, A. (2007) Pcons.net: protein structure prediction meta server. Nucleic Acids Res 35 (suppl_2) : W369-W374.
2. Lundstrom, J., Rychlewski, L., Bujnicki, J. and Elofsson, A. (2001) Pcons: a neural-network-based consensus predictor that improves fold recognition. Protein Sci 10 (11): 2354-2362
3. Wallner, B. and Elofsson, A. (2005) Pcons5: combining consensus, structural evaluation and fold recognition scores. Bioinformatics 21 (23): 4248-4254.
4. Cristobal, S., Zemla, A., Fischer, D., Rychlewski, L. and Elofsson, A. (2001) A study of quality measures for protein threading models. BMC Bioinformatics 2: 5.
5. Larsson, P., Wallner, B., Lindahl, E. and Elofsson, A. (2008) Using multiple templates to improve quality of homology models in automated homology modeling. Protein Sci 17 (6): 990-1002.
6. Wallner, B. and Elofsson, A. (2003) Can correct protein models be identified? Protein Sci 12 (5): 1073-1086
7. Ray, A and Wallner, B (submitted) ProQ2

FALCON-SWIFT

A threading method based on Short-cut phenomena

Mingfu Shao[1], Chao Wang[1], Sheng Wang[2], and Dongbo Bu[1]
[1] - *Institute of Computing Technology, Chinese Academic of Sciences*
[2] - *Institute of Theoretical Physics, Chinese Academic of Sciences*
dbu@ict.ac.edu

Our CASP9 server, FALCON-SWIFT, consists of a threading method (called SWIFT) and the Ab-initio method FALCON[1].

SWIFT is developed based on an interesting "short-cut" phenomena---in the optimal structural alignments between some CASP8 targets and templates in PDB, two spatially close cores of a template are aligned to continuous positions in query target, making the intermediate regions between the two aligned cores are totally "short-cut". Most of state-to-art threading methods perform badly in the "short-cut" cases since all cores are generally required to be aligned via a high gap penalty.

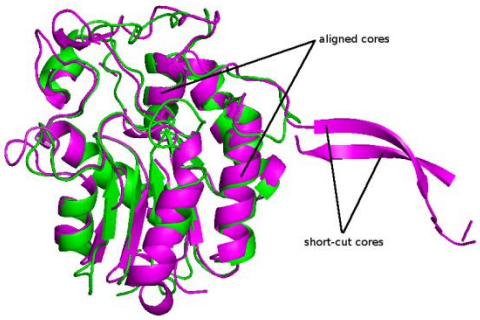


Figure: Short-cut phenomena illustration.
Structural alignment between native structure of CASP8 target T0448(green) and structure of 3BZWA(magenta) is shown. TMscore between them is 0.860. Two spatially close cores are aligned to continuous positions in query while cores between them are totally "short-cut".

FALCON employs a position-specific hidden Markov model to predict protein structure. The framework of FALCON can naturally repeat itself to converge to a final target, refining the decoy quality gradually.

Methods

SWIFT: We first give a formal definition for the "short-cut" phenomena, then design a new score function to characterize it. The scoring function is the linear weighted sum of several scoring items, including mutation, secondary structure, conformation letter similarity, solvent accessibility, environment and short-cut item. The short-cut item was designed specifically to allow and access short-cut. Even containing short-cut item in the scoring function, dynamic programming technique still apply though the time-complexity is increased.

FALCON: Cosine models are used to describe the local bias of a residue's torsion angle pair; a position-specific hidden Markov model(Fragment-HMM) is used to capture the dependencies among local biases of adjacent residues and sample a sequence of torsion angle pairs based on carefully selected fragments; an iterative strategy is used to increase the quality of the final decoys: the generated decoys are fed back to produce more accurate estimations of local structural biases and a more accurate Fragment-HMM.

Results

We conducted comparison of SWIFT against state-of-art threading methods on commonly used benchmarks: On four representative CASP8 targets with "short-cut" phenomena, our method can generate high-quality alignment while
RAPTOR[2] and HHpred[3] fail. On Prosup dataset, our method performs 6.3% better than RAPTOR in the measure of the alignment accuracy. In addition, our method show comparable fold recognition performance with other methods in the family level. We also evaluate the quality of the final predicted

structures. On 20 representative CASP8 targets, our method shows a comparable performance with Zhang-Server, and 0.02 better than RAPTOR, and 0.06 better than HHpred in the measure of average GDT_TS score.

Initial implementation of FALCON converges to within 6 A of the native for 100% of decoys on all six standard benchmark proteins used in ROSETTA[4] (discussed by Simons and colleagues in a recent paper), which achieved only 14%–94% for the same data. The qualities of the best decoys and the final decoys our theory converges to are also notably better.

Availability

Web-server can be accessed through http://www.bioinfo.org.cn/SWIFT/.

1. Li SC, Bu D, Xu J, Li M. (2008). Fragment-HMM: A new approach to protein structure prediction. *Protein Sci*. 2008 Nov;17(11):1925-34.
2. Xu J, Li M, Kim D, Xu Y. RAPTOR: optimal protein threading by linear programming. *J Bioinform Comput Biol*. 2003 Apr;1(1):95-117.
3. Söding J. Protein homology detection by HMM-HMM comparison. *Bioinformatics*. 2005 Apr 1;21(7):951-60. Epub 2004 Nov 5.
4. Simons KT, Kooperberg C,Baker D.Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol*. 1997 Apr 25;268 (1) :209-25.

FAMS-ACE3

Structure evaluation program using consensus method and circle QA program

Genki Terashi[1], Kazuhiko Kanou[1,2], Makoto Oosawa[1], Yuuki Nakamura[1], Hideaki Umeyama[1], and Mayuko Takeda-Shitaka[1]
[1] - *School of Pharmacy, Kitasato University*
2- Infectious Disease Surveillance Center, National Institute of Infectious Disease
terashig@pharm.kitasato-u.ac.jp

In the CASP9, our fams-ace3 server participated in the TS prediction category as a human expert group. We applied two different type of scoring functions for the fully automated model prediction server, fams-ace3: (1) the local and global consensus score; and (2) the model quality score based on classification of the side-chain environment for each residue. The consensus methods were used as a filter to select the models which have high structural conservation comparing with the set of models. The fams-ace3 differs from previous procedure (fams-ace2) in the step of consensus method. We introduced a global consensus method and a variation value of server models. The model quality score was used for the final selection of the best model. This model quality score was calculated by our model quality assessment program CIRCLE[1].

Methods

The procedure of fams-ace3 can be summarized as the following 4 steps: (1) incorrect models which have serious physical clashes or broken main-chain structures were removed. (2) Superposition of the server models were carried out to calculate the variation values (*Var*) of targets. The variation value is:

$$Var = \frac{\sum\sum SIM_{3.5}(i,j)}{\sum\sum COM(i,j)} \quad (1)$$

Where $SIM_{3.5}(i,j)$ was the number of structurally aligned residues which superimpose well (within 3.5Å) between model *i* and *j*. $COM(i,j)$ was the number of common residues from model *i* and *j*. According to the variation value, we define EASY(*Var*>=0.7) and HARD(*Var* <0.7) targets. (3) The consensus scores were calculated as follows:

$$Cons_m \begin{cases} \dfrac{\sum_{n}^{N} SIM_{3.5}(m,n)}{N} & EASY \\ \dfrac{\sum_{n}^{N}\sum_{i}^{R_m} SIM_{3.0}(LOC_{m,j}, LOC_{n,i})}{N} & HARD \end{cases}$$

Where *N* is the number of server models. *LOCm,i* is a subset of C-alpha coordinates which exist within 10Å from the *i* th residue of model *m*. $SIM_r(a,b)$ is a maximum number of C-alpha coordinates (subset *a*) which superimpose well (within *r*Å) upon their corresponding C-alpha coordinates in subset b. The top 20% and 10% of server models were selected in the order of the consensus score, for EASY and HARD targets, respectively. (4) All of the server models, selected in step (3), were refined and rebuilt utilizing our homology modeling program FAMS[2]. (5) The top 5 structures were selected, according to a model quality evaluation based on their CIRCLE score. The fams-ace3 is a fully automated server and does not require human intervention. The parameters of fams-ace3 were optimized by the data set of previous CASP8.

Results

Our automated evaluations on available 85 targets of CASP9 are:

Total GDT_TS	4863.70
Average GDT_TS	57.22
Number of correct side-chain (chi1)	5177

The fams-ace3 performed well in comparison with other server models.

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

FAMSD

Individual comparative modeling server using FAMS-MULTI, CIRCLE and SPLICER.

Kazuhiko Kanou[1,2], Genki Terashi[1], Yuuki Nakamura[1], Makoto Oosawa[1], Hideaki Umeyama[1] and Mayuko Takeda-Shitaka[1]

[1] - School of Pharmacy, Kitasato University
2- Infectious Disease Surveillance Center, National Institute of Infectious Disease
kanouk@pharm.kitasato-u.ac.jp

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

METHOD

(1) Making sequence alignments

10 kinds of alignment programs, BLAST, PSI-BLAST[4], PSF-BLAST, RPS-BLAST, IMPALA, Pfam-BLAST, CSI-BLAST, SPARKS2[5], SP3[6] and HHsearch[1] were executed for each target protein sequence. Various alignments were generated and were filtered with its alignment score. The alignment scores for 7 kinds of BLAST related alignment methods were calculated with following equation,

$$Score_{ali} = k_i \times Len \times SEQid^{m} \times SS^{n} \quad (1)$$

Here *Len* is the number of residues of a predicted model. *Hom* indicates sequence identity % value, *SS* is the degree of secondary structure agreement between the secondary structures predicted one from sequence using PSI-PRED[7] and one calculated from model using STRIDE. k_i is a coefficients for each alignment method. The ki value and the parameters (m, n) are optimized using CASP5 target proteins for each sequence identity level[8].
And as the alignment score for SPARKS2 and SP3, Z-score of their output was used. When the alignment score was more than (the maximum score of all alignments) * X, these alignments were used to construct model. A parameter X is a cut-off value which was decided using CASP7 targets as a training set depending on difficulty of each target[9]. The difficulty was predicted using Support Vector Machine (SVM). The alignment score and sequence identity of PSI-BLAST and these of SPARKS are used as parameters for SVM training.
As the alignment score for HHsearch, score of its output was used. Then value of X was set to 0.9 which was determined by training with CASP8 targets.

(2) Constructing three-dimensional structures

We constructed 3D structures using FAMS program based on each selected alignment which was mentioned in the preceding section.

(3) Selecting the best structure

In the FAMSD, SPLICER method was used in the 3D model selection (The details of SPLICER method are mentioned in the abstract of "Splicer"). Six evaluation scores for a model were calculated for each constructed models. Six scores were CIRCLE score, *SSscore, Vhp, Vhb, Vcoli* and *Vrama*. From these evaluation scores and model length, the predicited GDT_TS value was obtained using the non-linear regression method. The predicted GDT_TS., called *sGDT_TS*, was calculated with R program[10] using the gam(Generalized Additive Models) function.

$$sGDT\_TS = pred\,[gam\{s(CIRCLE) + s(SSconf) + s(V_{HP}) + s(V_{HB}) + s(V_{coli}) + s(V_{rama}) + s(len_{mdl})\}]/(4*\text{target length}) \quad (2)$$

The model which had max *sGDT_TS* value among the constructed models was selected as a best structure for the target protein.

(4) Refinement of the selected models

The model selected in the previous step was reconstructed using FAMS-MULTI program as an alternative of FAMS based on the same alignment. The FAMS-MULTI uses multiple template proteins and the correctness of FAMS-MULTI models is superior to FAMS model. Furthermore, the reconstructed model by FAMS-MULTI was refined by using Molecular Mechanics & Molecular Dynamics refinement program. With this refinement procedure, hydrogen bonds, main chain torsion angles and side-chain torsion angles were refined slightly and collisions of hydrophobic atoms were decreased[9].

RESULTS

Figure 1 shows the distribution of alignment method of finally ranked first models by our selecting method mentioned above.

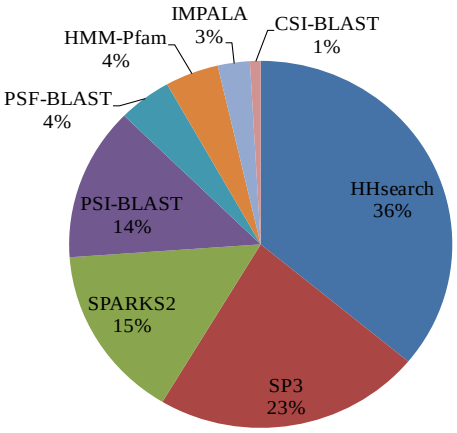


Figure 1

We implemented automated evaluations using 85 experimental structures of 129 CASP9 targets became available by September 12, 2010. Table 1 shows the summary of the results for FAMSD_TS1.

Total GDT_TS	4567.94
Average GDT_TS	53.74
Number of correct side-chain (chi1)	4769

Table 1.

1. Söding J., Bioinformatics. 2005 Apr 1;21(7):951-60
2. Ogata, K. and Umeyama, H. J Mol Graph Model 2000; 18, 258-272.
3. Terashi G, Takeda-Shitaka M, Kanou K, Iwadate M, Takaya D, Hosoi A, Ohta K, Umeyama H. Proteins. 2007;69 Suppl 8:98-107.
4. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. Nucleic Acids Res. 1997; 25, 3389-3402.
5. Zhou H., Zhou Y., Proteins, 55(4), 1005–1013 (2004).
6. Zhou H., Zhou Y., Proteins, 58(2), 321–328 (2005).
7. Jones DT., J Mol Biol. 1999 Sep 17;292(2):195-202.
8. Iwadate M, Kanou K, Terashi G, Umeyama H, Takeda-Shitaka M. Chem Pharm Bull (Tokyo). 2010 Jan;58(1):1-10.
9. Kanou K, Iwadate M, Hirata T, Terashi G, Umeyama H, Takeda-Shitaka M. Chem Pharm Bull (Tokyo). 2009 Dec;57(12):1335-42.
10. The R project homepage : http://www.r-project.org/

FAMS-MULTI

FAMS-MULTI: An automated homology modeling based upon multiple reference proteins using better pairwise alignments

Kazuhiko Kanou[1,2], Genki Terashi[1], Yuuki Nakamura[1], Makoto Oosawa[1], Hideaki Umeyama[1] and Mayuko Takeda-Shitaka[1]

[1] - School of Pharmacy, Kitasato University
2- Infectious Disease Surveillance Center, National Institute of Infectious Disease
kanouk@pharm.kitasato-u.ac.jp

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

METHODS

1. Generation of better pairwise alignments

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

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

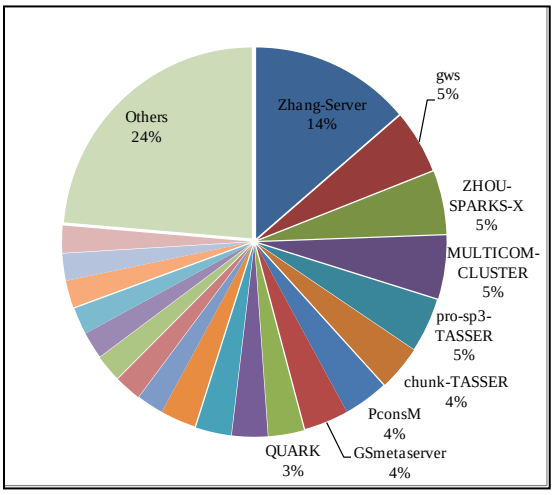


Figure 1

2. Construction of models by FAMS-MULTI

First, template protein was divided based on domain definition of SCOP [4]. Some reference proteins were chosen based on the sequence and structural similarity with each template domain. Next, a multiple structural alignment based on the superposition of Cα atoms was performed among the reference proteins including in the template. The target sequence was put on for this alignment based on the piarwise alignment between target and template mentioned in the preceding section. Thus, we get a result of multiple alignments between a target protein and reference proteins.

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

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

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

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

3. Evaluate models

Thus, some full atom models were constructed. These models were evaluated and selected based on the fams-ace2 selecting method (combined Cα consensus and Circle score [5]). Consequently top five models were selected.

4. Refine models

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

Overall procedure of fams-multi was shown as in Figure. 1. All procedures of human expert team fams-multi were implemented fully automatically.

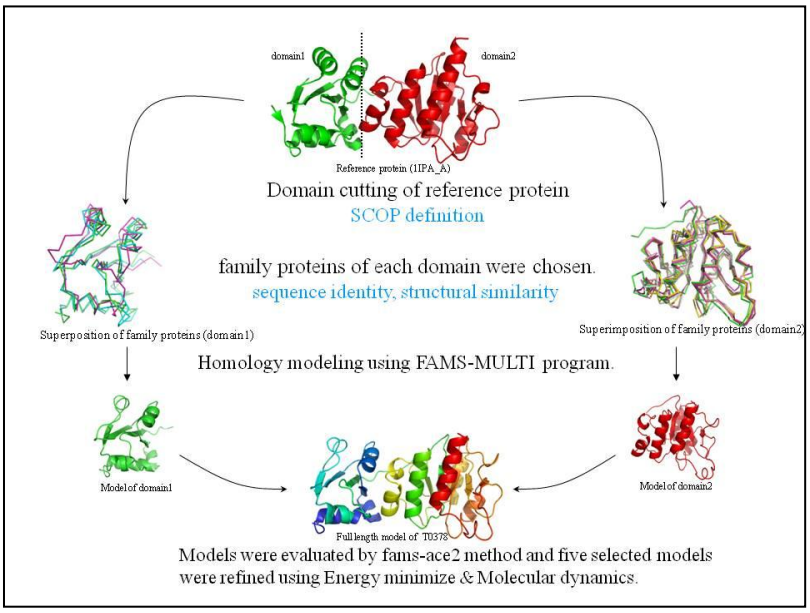


Figure 1. Overall procedure of fams-multi.

RESULTS & DISCUSSION

We implemented automated evaluations using 85 experimental structures of 129 CASP9 targets became available by September 12, 2010. Table 1 shows the summary of the results for fams-multi_TS1.

Total GDT_TS	4833.18
Average GDT_TS	56.86
Number of correct side-chain (χ1)	5038

1. Ogata, K. and Umeyama, H. J Mol Graph Model 2000; 18, 258-272.
2. Ogata K, Umeyama H. Proteins. 1998; 31(4):355-69.
3. Shindyalov IN, Bourne PE. Protein Engineering 1998; 11(9) 739-747.
4. Andreeva A., Howorth D., Chandonia J.-M., Brenner S.E., Hubbard T.J.P., Chothia C., Murzin A.G. Nucl. Acids Res. 2008 36: D419-D425.
5. Terashi G, Takeda-Shitaka M, Kanou K, Iwadate M, Takaya D, Hosoi A, Ohta K, Umeyama H. Proteins. 2007;69 Suppl 8:98-107.

FAMSSEC

Model selection method based on the side chain environment consensus score

Hideaki Umeyama[1,2], Katsuichiro Komatsu[1], Kazuhiko Kanou [3], Genki Terashi [1],
Mayuko Takeda-Shitaka [1,2]
Kitasato univ.[1], Riken[2], NIID Japan[4]
umeyamah@pharm.kitasato-u.ac.jp

A consensus method like 3D-Jury [1] is one of the most powerful methods of model quality assessment. 3D-Jury score represent consensus of the backbone geometry among structure models. This method can select "good backbone" models but the quality of the side chain of selected models is not so good.
Thus we developed a new consensus method which considers side chain environment for the purpose of selecting good side chain models [2]. This method should be able to select good side-chain models among many server models. We participated in TS category using this method as a human expert team FAMSSEC. We describe the algorithm of this method.

METHODS
First, we calculated the side chain environment composed of 'fraction buried' and 'fraction polar' for each residue of predicted model. 'Fraction buried' is the fraction of buried area within the surrounding side chain atoms, and 'fraction polar' is the fraction of buried area within the surrounding polar atoms. These values range from 0 to 1.0 per residue. When the model A was assessed, for each residue of model A, the side chain environment was calculated and is compared with the other models. If the Euclidian distance between the side chain environment ('fraction buried' and 'fraction polar') [3] of one residue of model A and that of corresponding residue of another model was within 0.2, we considered that the two residues were in the same environment. For each model, we counted the number of residues in the same environment and the side chain environment score is the summation of those numbers. The threshold of 0.2 was determined using CASP7 models as a training set.

In CASP8, we participated in QA category as a team 'FAMSD_QA'. We had refined all predicted models by FAMS [4] and had assessed quality of these models using following combined score.

$$\text{score} = env\_con + w * SSscore$$

Here, *env_con* represents the side chain environment consensus score and *SSscore* represents the degree of match between the secondary structure of a predicted model and the secondary structure predicted from the given sequence with PSIPRED [5]. *w* is the weighting factor for *SSscore* and ranges from 0 to 1 depending on the predicted difficulty using SVM (Table 1). In the case of difficult targets, more weight is given to *SSscore* than easy targets. This value was optimized using CASP7 models.

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

Table 1. Values of *w*.
PSIB and SPK2 indicate the predicted difficulty using alignment score and sequence identity of PSI-BLAST and these of SPARKS2, respectively, as parameters for SVM training.

As the result of the verification using the CASP8 server models, it was found that the SEC method selects the models with more accurate positioning of the side-chain atoms than the 3D-Jury method. Thus the SEC method was used in combination with the 3D-Jury method (3DJ+SEC) so that models were selected with improved quality both in the CA backbone and side-chain atom positions.
Moreover, the CIRCLE (CCL) method based on the 3D-1D profile score has been shown to select the best possible models that are the closest to the native structures from candidate models. Accordingly, the 3DJ+SEC+CCL method, in which CIRCLE is used after reducing the number of candidates by the 3DJ+SEC consensus method, was found to be very effective in selecting high quality models.

1. Ginalski K, Elofsson A, Fischer D, Rychlewski L. Bioinformatics. 2003; 19:1015–8
2. Kanou K, Hirata T, Terashi G, Umeyama H, Takeda-Shitaka M. Chem Pharm Bull (Tokyo). 2010 Feb;58(2):180-90.
3. Terashi G, Takeda-Shitaka M, Kanou K, Iwadate M, Takaya D, Hosoi A, Ohta K, Umeyama H. Proteins. 2007;69 Suppl 8:98-107.
4. Ogata, K. and Umeyama, H. An automatic homology modeling method consisting of database searches and simulated annealing J Mol Graph Model. 2000; 18, 258-272
5. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 1999; 292: 195-202.

FAMSSEC

FAMS modeling of complex proteins and prediction of ligand binding sites by integrated-FAMSD

Hideaki Umeyama[1,2], Katsuichiro Komatsu[1], Kazuhiko Kano [3], Genki Terashi [1], Mitsuo Iwadate [2,4], Mayuko Takeda-Shitaka [1,2]
Kitasato univ.[1], Riken[2], NIID Japan[3], Chuo univ. [4]
umeyamah@pharm.kitasato-u.ac.jp

We have developed the modeling system for the complex protein in addition to the isolated protein. Our method is based on the homology modeling, in which the Monte Carlo method is applied in the construction of the main chain and in the concept of local space homology. Our modeling program called FAMS[1] is active in relation to the CASP contests during these ten years, and it has more powerful activity in adding the CIRCLE[2] program, which estimates energetically the protein structure after the protein folding from a free energy point of view.

Methods

By using this system, we participated in the function prediction in which we predict the binding site of some ligands such as substrates, inhibitors, antagonists or agonists. First, tens of sequence alignments were obtained in the descending order of the predicted GDT_TS using the power function program[3] based on several blast programs. Second, whether each of clustered reference proteins similar to the query protein included low molecular weight compounds or not was manually observed. When we looked for small ligands, and ions except for sodium and chlorine such as Ca, Zn, and Cd in modeling structures were also treated as the ligand. We ignored compounds like DNA, RNA and large proteins. We executed these FAMSD[4] calculations using Web browser aided in those on stand alone Linix PC with Intel Core2 Quad CPU, 1.5TB Hard Disk, and 2GB Memory, whose the system is called "integrated-FAMSD".

(1) Reference proteins including ligands are superimposed to the top rank model, and we searched whether the superimposed ligands overlap each other or not. The places indicates by the overlap was treated as ligand binding site. After that, we searched that some ligands can interact with the target protein in no short contacts. Moreover, the hydrogen bonds and hydrophobic interaction were checked.

(2) When there were no ligand binding sites for the top rank model, we searched the site with the same procedure as (1) in order of rank.
(3) When common metal ions are included in some reference proteins, we superimposed some reference proteins into the model and searched the position of each metal ion. When the position is near each other, and each ion coordinated to two or more amino acid residues which have the lone pair around the binding site, it is estimated as the ligand.

Results

128 targets of T0515-T0643 except for T0637 canceled with human were submitted to the prediction center. We could predict 78 binding sites in 128 targets, but could not do 50 ones. There were no binding sites in 50 targets.

Availability

Anyone can easily construct FAMS models including ligands using the Web interface on integrated-FAMSD. If the model is a receptor for a drug, it can be used as a binding site of ligand for insilico screening.

1. Ohta, K. & Umeyama, H. (2010). An automatic homology modeling method consisting of database searches and simulated anneannealing. *J. Mol. Graph. Molel.,* **18**, 252-272 & 305-306.
2. Terashi, G., Takeda-Shitaka, M., Kanou, K., Iwadate, M., Takaya, D., Hosoi, A., Ohta, K. & Umeyama, H. (2007). Fams-ace: a combined method to select the best model after remodeling all server models. *Proteins*, **69** Suppl 8, 98-107.
3. Iwadate, M., Kanou, K., Terashi, G., Umeyama, H. & Takeda-Shitaka, M. (2010). Method for predicting homology modeling accuracy from amino acid sequence alignment: the power function. *Chem. Pharm. Bull.*, **58**, 1-10.
4. Kanou, K., Iwadate, M., Hirata, T., Terashi, G., Umeyama, H. & Takeda-Shitaka, M. (2009). FAMSD: A powerful protein modeling platform that combines alignment methods, homology modeling, 3D structure quality estimation and molecular dynamics. *Chem. Pharm. Bull.*, **57**, 1335-1342.

FEIG

See PRECORS, PRECORS-QA

FFAS03
FFAS03n
FFAS03ss
FFAS03a

VERSIONS OF FFAS METHOD TESTED IN CASP9 EXPERIMENT

Zhanwen Li, Lukasz Jaroszewski, Christian Zmasek, and Adam Godzik
Sanford Burnham Medical Research Institute
adam@burnham.org

In the CASP09 experiment we tested four versions of the FFAS profile-profile alignment algorithm, including three newly developed variants:

- FFAS03 - the current version of the FFAS algorithm as implemented on the public ffas server at ffas.burnham.org.
- FFAS03n - newly re-optimized ffas algorithm, as tested on a large benchmark for recognition of remote similarities. This version uses a new substitution matrix proposed by Price et al. (Bioinformatics. 2005 Dec 1;21(23):4318), includes counts of gaps in profile-profile alignment and gap penalties optimized specifically for the recognition of very remote homologs (different superfamilies in SCOP database).
- FFAS03ss - similar to FFAS03n but it also includes secondary structure matching score.
- FFAS03a - a new version of the FFAS03 algorithm with profiles seeded with ancestral sequences reconstructed for a protein family. The ancestral root sequence was reconstructed by the ANCESCON program (http://prodata.swmed.edu/ancescon/ancescon.php) from the multiple sequence alignment built by MAFFT program (http://mafft.cbrc.jp/alignment/software/).

Maintenance of the FFAS server is supported by the grant R01-GM087218-01 from the National Institute of General Medical Sciences.

Firestar
CNIO-Firestar

Server and human predictions for firestar

P. Maietta[1], A. del Pozo[1], M.L. Tress[1]* and G. Lopez[1]*
[1]-CNIO *(Spanish National Cancer Research Centre), Madrid, Spain*
glopez@cnio.es, mtress@cnio.es

Here we describe the protocols for obtaining and scoring ligand binding predictions and template based 3D models for the 9th edition of CASP. The main motivation of our group was to determine to what extent structural information could be used to improve ligand binding predictions, while we also investigated whether functional information might be used to obtain better 3D models.

Methods

firestar (1) is an expert system for predicting ligand binding and catalytic sites. Predictions are based on the large catalogue of sites collected from PDB structures in the FireDB (2) resource. The *firestar* server required human intervention, but for CASP9 we developed a version of *firestar* that works automatically. This new version also avoids over-prediction of residues and was designed to maximize the MCC scoring criteria from the previous edition of the CASP ligand binding experiment (3).

For the "human" method we obtained additional functional information with SIAM (unpublished) in the form of Gene Ontology terms. This method helped us to identify probable cognate ligand specificity binding for the targets. In a second step we used structural information to modify the automatic *firestar* predictions. We generated multiple models for target domains using a range of profile-based search methods and MODELLER (4). We chose the best scoring models, based on coverage and global and local scores for the quality of alignments from SQUARE (4), for the refinement of the *firestar* predictions.

The models were superposed onto structures that bound the *firestar* predicted ligands using LGA (6). After mapping the predicted functional residues onto the structure, we calculated distances between putative ligands and the surrounding amino acids. A final decision for each predicted residue was made by taking into account distances to the superimposed ligand, conservation of the residues in the model and SQUARE reliability scores.

Results

The *firestar* server returned binding site predictions for 60 targets, 19 were metal binding sites, 17 nucleotide and cofactor sites and 21 sites for metabolites. Three cases were predicted to bind SO4 and these probably probably mimic the cognate ligand. The CNIO-firestar predictor modified the list of residues predicted by *firestar* for all but 10 targets. We also added two predictions where *firestar* alone could not make a reliable enough prediction.

46 of the first models submitted by the CNIO-firestar group for the 3D structure prediction experiment were used in the prediction of binding residues by the CNIO-firestar group because their binding sites were conserved between targets and templates.

Availability

Our servers *firestar,* FireDB and SQUARE can be accessed via http at http://wwwfiredb.bioinfo.cnio.es

Our modelling pipeline is a prototype and will be accessible in the future. We hope to integrate a structural module in *firestar*.

1. Lopez, G, Valencia, A and Tress, ML. (2007). *firestar* - Prediction of functionally important residues using structural templates and alignment reliability. *Nucleic Acids Res.* **35** W573;
2. Lopez, G, Valencia, A and Tress, ML. (2007). FireDB - a database of functionally important residues from proteins of known structure. *Nucleic Acids Res,* **35**, D219;
3. Lopez, G., Ezkurdia, I., Tress ML. *Assessment of ligand binding residue predictions in CASP8.* Proteins. 2009 Jul 22. PMID: 19714771
4. Fiser, A., Sali, A. (2003). Modeller: generation and refinement of homology-based protein structure models. *Methods Enzymol.* 374, 461;
5. Tress, ML, Jones, DT and Valencia, A. (2003). Predicting reliable regions in protein alignments from sequence profiles. *J Mol Biol*, 330, 705;
6. Zemla, A. (2003). LGA: A method for finding 3D similarities in protein structures. *Nucleic Acid Res*, **31**, 3370.

FLOUDAS

Enhanced Astro-Fold for three dimensional structure prediction of proteins: A first principles approach

Y. Wei, A. Subramani and C. A. Floudas
Department of Chemical and Biological Engineering, Princeton University
floudas@titan.princeton.edu

Astro-Fold[1,2], a first principles method for protein structure prediction, is based on an overall deterministic global optimization framework (αBB) coupled with a stochastic algorithm, conformational space annealing (CSA). It consists of several steps including secondary structure prediction, residue contact prediction, loop prediction, hybrid method for structure generation, near-native structure identification, and chemical shift-based structure refinement.

Methods

The first stage of Astro-Fold predicts helical and beta-sheet structures[3,4]. A consensus method for secondary structure prediction has been developed based on seven prediction methods[3]. The consensus method is a MILP (Mixed Integer Linear Programming) model that maximizes the number of correctly predicted amino acids for a training set of proteins. The prediction of beta-sheet and disulfide bridge topology is based on an ILP (Integer Linear Programming) model in which the hydrophobic contact energy between strands is maximized to derive the optimal topology[4]. A number of additional constraints are added to obtain biologically relevant topologies.

The second stage predicts angle and distance restraints through residue contact prediction and loop prediction. The residue contact prediction is based on a novel ILP model that predicts contacts by minimizing the total statistical energy of a protein subject to a set of physically observed constraints[5]. Restraints are determined for the loop residues connecting helical and strand regions through an iterative formulation involving dihedral angle sampling, constrained nonlinear optimization of ECEPP/3 force field, and a novel clustering approach[6].

Based on the constraints predicted from the previous stages, a hybrid algorithm that combines the deterministic αBB global optimization algorithm[7], stochastic global optimization (CSA)[8], and molecular dynamics in torsion-angle space is implemented to solve the constrained non-convex global optimization problem[1,2]. The features of αBB provide valid lower bounds and a theoretical guarantee of convergence to the global optimum while the features of CSA provide upper bounds through extensive sampling of the energy landscape. This process relies on detailed atomistic modeling, constrained nonlinear optimization and a quick pre-cursor rotamer optimization to eliminate steric clashes[9].

At this stage, ICON, a novel iterative traveling-salesman problem-based clustering method, is used to identify the near-native structures of the protein. The iterative feature of ICON eliminates clusters of structures at each iteration based on a statistical analysis of cluster density and average spherical radius[10].

The selected structures are subject to chemical-shift-based structure prediction process. The chemical shifts are predicted through ShiftX and used by locally installed CS23D to re-predict the

structures[11,12]. These structures are used to generate tighter angle and distance constraints for a second iteration of tertiary structure prediction to generate the final set of predicted structures.

1. Klepeis J.L. and Floudas, C.A. (2003) ASTRO-FOLD: A combinatorial and global optimization framework for Ab initio prediction of three dimensional structures of proteins from the amino acid sequence. *Biophys. J.*, **85**, 2119-2146.
2. Klepeis J.L., Pieja M.J. and Floudas C.A. (2003) Hybrid global optimization algorithms for protein structure prediction: alternating hybrids. *Biophys. J. 84*, 869 – 882.
3. Wei Y. and Floudas C.A (2010) A consensus method for secondary structure prediction, In preparation.
4. Subramani A. and Floudas C.A. (2010) In preparation.
5. Rajgaria R., Wei Y. and Floudas C.A. (2010) Contact prediction for beta and alpha/beta proteins using integer linear optimization and its impact on the first principles 3D structure prediction method ASTRO-FOLD, *Proteins*, **78**, 1825-1846.
6. Subramani, A. and Floudas C.A. (2010) In preparation.
7. Androulakis I.P., Maranas C.D. and Floudas C.A. (1995) A global optimization method for general constrained nonconvex problems, *Journal of Global Optimization*, **7**,337-363.
8. Lee J., Scheraga H.A., and Rachovsky S. (1997) New optimization method for conformational energy calculations on polypeptides: conformational space annealing, *Journal of Computational Chemistry* **18**, 1222-32.
9. McAllister S.R. and Floudas C. A. (2010) An improved hybrid global optimization method for protein tertiary structure prediction, *Computational Optimization and Applications*, **45**, 377-413.
10. Subramani A., DiMaggio P.A. and Floudas C.A., (2009), Selecting high quality protein structures from diverse conformational ensembles, *Biophysical Journal*, **97**, 1728-1736.
11. Neal S., Nip A.M., Zhang H.Y. and Wishart D.S., (2003) Rapid and accurate calculation of protein ^{1}H, ^{13}C and ^{15}N chemical shifts, *Journal of Biomolecular NMR*, **25**,215-240.
12. Wishat D.S., Arndt D., Berjanskii M., Tang P., Zhou J. and Lin G. (2008) CS23D: a web server for rapid protein structure generation using NMR chemical shifts and sequence data. *Nucleic Acids Res*., **36**,w496-502.

FLyPred

Residue-residue contact prediction through matching of known motifs

F. Lysholm [1,2], P.Björkholm [3], T.R. Hvidsten [4], B. Persson [1,2]
[1] IFM Bioinformatics and Swedish e-Science Research Centre (SeRC), Linköping University, S-581 83 Linköping, Sweden, [2] Karolinska Institutet, Department of Cell and Molecular Biology, SE-17176 Stockholm, Sweden, [3]Center for Biomembrane Research, Department of Biochemistry and Biophysics, Stockholm University, SE-106 91 Stockholm, Sweden, [4] Umeå Plant Science Centre, Department of Plant Physiology, Umeå University, SE-901 87 Umeå, Sweden

Accurate residue-residue contact prediction in "template-free" targets would provide a huge leap forward for *ab initio* protein structure prediction. While short-range contact predictions are usually more accurately predicted than mid- and long-range contacts, they unfortunately also provide less new information. In contrast, accurate mid/long-range predictions minimize the search space or provide restrictions which would greatly aid *ab initio* methods. Therefore, residue-residue contact predictions of especially long-range contacts are an important field of bioinformatics (1) that could increase the number of targets for which a sufficiently accurate structure could be predicted (2).

Methods

Through the use of the ASTRAL v1.75 (3) subset of PDB filtered at 40% sequence identity, a novel method for identifying structural motifs was applied. Any two fragments of four amino-acids satisfying a minimum structural proximity threshold were considered a seed motif. By evaluating all possible combinations of such fragment-pairs (less than L^2 combinations per sequence, where L is the amino-acid sequence length) all possible seed motifs was found. For each such seed-motif the fragments were expanded through maximizing a motif score function. The score function was designed to maximize the number of interactions between the residues of the fragments, while assigning penalty to an increased sequence length. Finally, a set of non-redundant two fragment motifs was retained as well as all instances of three fragment motifs constructed from fragment pairs sharing only one fragment.

Each structural motif was assessed in terms of their conserved sequence profile, secondary structure, fragment separation and amino-acid composition. A sequence profile (PSSM) using 3 iterations of PSI-BLAST (4) against NCBI NR and secondary structure information predicted by PSIPRED (5) was pre-calculated for each ASTRAL entry. Using this information, a model was trained to recognize the motifs in the training set using 10-fold cross-validation. All motifs which were distinguishable by sequence signal (at least 20% accuracy during cross-validation) were stored in a motif database. Furthermore, the estimated predictive power of each motif when re-queried against the training set was calculated along with information of structural deviation within the motifs.

For a query sequence, a PSSM and the secondary structure were calculated before the motif database is searched. Local structure, predicted by the matched motifs, can then be assigned to the query sequence. All residue-residue distances are extracted for matched motifs and compared resulting in a list of predicted residue-residue contacts < 8 Å. Contacts are ordered by descending estimated p-value and for each type of contact (short-, medium- and long-range) up-to $N = Pct \cdot L$ contacts are presented, where L is the sequence length and Pct is a real number (default 0.5).

A consensus server, FragFly, was also construcuted where the results of FLyPred and FragHMMent (6) are combined, weighting each method equally.

Availability

A prediction server is available at http://bioinfo8.limbo.ifm.liu.se/FLyPred/ and at http://bioinfo8.limbo.ifm.liu.se/FragFly/

1. Wu, Sitao and Zhang, Yang. (2008). A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. Bioinformatics, Vol. 24, pp. 924-931.
2. Floudas, C.A, et al., et al. (2006). Advances in protein structure prediction and de novo protein design: A review. Chemical Engineering Science. 61, pp. 966 – 988.
3. Chandonia, John-Marc, et al. (2004). The ASTRAL Compendium in 2004. Nucleic Acids Res, Vol. 32, pp. D189--D192.
4. Altschul, S. F., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res, Vol. 25, pp. 3389-3402.
5. Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol, Vol. 292, pp. 195-202.
6. Björkholm, P., et al. (2009). Using multi-data hidden Markov models trained on local neighborhoods of protein structure to predict residue-residue contactsBioinformatics, Vol. 25, pp. 1264-1270.

FOLDIT

Multiplayer online game-based homology and ab-initio modeling

F. Khatib[1], S. Cooper[1], J. Thompson[1], I. Makedon[1], J. Barbero[1], Z. Popović[1],
D. Baker[1] and Foldit players[2]
[1] - University of Washington, Seattle, WA, [2] - Worldwide
dabaker@uw.edu

Models were constructed using Foldit, the online multiplayer game at http://fold.it. CASP9 targets shorter than 165 residues and not designated as multimers by the CASP organizers were given to Foldit players as puzzles to solve.

Methods

Foldit uses the Rosetta protein modeling software package[1] and allows players to modify and visualize protein structures in real time[2]. Foldit players are provided with tools that allow them to directly move the protein structure manually, such as directly pulling on any part of the protein. They are also able to rotate helices and rewire beta-sheet connectivity. Players are able to guide moves by introducing soft constraints and fixing degrees of freedom, and have the ability to change the strength of the repulsion term to allow more freedom of movement. Available automatic moves–combinatorial side-chain rotamer packing, gradient-based minimization, fragment insertion–are Rosetta optimizations modified to suit direct protein interaction and simplified to run at interactive speeds. Each CASP9 puzzle was typically accessible to Foldit players for 7-8 days.

For de novo targets, models were constructed using the five BAKER-ROSETTASERVER predictions. Foldit players were given each BAKER-ROSETTASERVER model as a puzzle to refine. For comparative modeling targets, Foldit puzzles started from an extended chain, with alignments to known templates from the HHsearch server provided[3]. Foldit players were able to modify alignments between the query and template sequences within the game. They could then build models based on these alignments by threading the query sequence onto the templates and refining these models using the tools listed above. CASP9 targets with terminal regions highly predicted to be disordered, according to the metaPrDOS server[4], were trimmed before being given to Foldit players.

Quality and ranking of individual models was determined entirely by the Rosetta full-atom energy. A conformationally diverse set of Foldit submissions were selected from the top-ranking Foldit predictions.

Availability

Foldit is available through the Rosetta Commons at www.rosettacommons.org.

1. Leaver-Fay,A., Tyka,M., Lewis,S., Lange,O.F., Thompson,J., Jacak,R., Kaufman,K., Renfrew,P.D., Smith,C., Sheffler,W., Davis,I., Cooper,S., Treuille,A., Mandell,D., Richter,F., Ban,Y.A., Fleishman,S., Corn,J., Kim,D.E., Lyskov,S., Berrondo,M., Mentzer,S., Popović,Z., Havranek,J., Karanicolas,J., Das,R., Meiler,J., Kortemme,T., Gray,J.J., Kuhlman,B., Baker,D. & Bradley,P. (2010) ROSETTA3.0: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules. *Meth Enz.* In Press.
2. Cooper,S., Khatib,F., Treuille,A., Barbero,J., Lee,J., Beenen,M., Leaver-Fay,A., Baker,D., Popović,Z. & Foldit Players (2010) Predicting protein structures with a multiplayer online game. *Nature*. ***466***, 756-760.
3. Söding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics.* **21**(7):951-60.
4. Ishida,T. & Kinoshita,K. (2008) Prediction of disordered regions in proteins based on the meta approach. *Bioinformatics*, **24**(11):1344-1348.

Formann_Server

Fast algorithm with template-free based modeling

Yeona Kang[1], Charles M. Fortmann[1]
[1] –*Material Science Department at Stony Brook University*
yeona@ams.sunysb.edu

In the ninth Critical Assessment of techniques for protein Structure Prediction (CASP9), we participated as FKinitial in the "Human-Server" category and, Fortmann_server in the "Server-Only" category. Fortmann's group used a combination of original physical secondary and tertiary structure prediction methods for 3D structure prediction running on desktop computers. These physical models employed no statistical templates (protein structure determination without appeal to templates has long been sought). This work describes the fundamental physical model, the algorithms, and accuracy. For example, the model considers the diffusion and drift of charged and un-charged protein regions relative to one another in multiple energy-type field gradients (forces). Where forces including: electrostatic, hydrophobic and steric hinderance is described by a single unified force expression based upon physical parameter inputs. The physical underpinnings of this algorithm provide insight into the mechanisms of secondary and tertiary structure generation and pave the pathway to faster all *ab-initio* structure prediction.

Methods

First a fast secondary structure algorithm running on a desktop computer is applied to locate and tag all secondary structures within the target proteins. After running secondary structure determination program, the secondary structure output file identifies all alpha helix regions and most beta sheet regions. The initial file is reconfigured PDB format in some cases with assistance from Pymol software. The resultant PDB file (with identified secondary structures) was used as input file of tertiary structure prediction. This operation consumes less than one minute CPU time.

Next tertiary structure is determined by applying a second all original algorithm that tracks drift and diffusion of the protein relative to itself in the various force fields self-generated by the protein. The interaction between two residues (e.g., resulting from a charge-charge interaction) was simplified by taking the center of interaction to reside on the appropriate alpha carbon atoms. Both motion in the various force fields (drift) and the thermal motion (diffusion) were considered. The aforementioned multiple field considerations were used for the core drift calculation. The simulations were run assuming laboratory temperature (300 K). Running of a desktop computer this second algorithm required a few minutes CPU time. Since the original model can quickly and accurately track early time folding events even on large proteins, standard protein folding software could be applied to the resultant structures for greater accuracy without a large time penalty. Here AMBER-based energy minimization (relaxation) steps were applied to structures generated by the second algorithm for the better accuracy.

1. Yeona Kang, Enrique Jean, and C.M. Fortmann. (2006) Einstein relations for energy coupled particle systems Appl. Phys. Lett. 88, 11, 112110-1-112110-3
2. Yeona Kang and C.M. Fortmann. (2007) A structural basis for the Hodgkin and Huxley relation Appl. Phys. Lett. 91, 22, 223903-1-223903-3

FragHMMent

FragHMMent – Contact prediction using hidden Markov models trained on alignments of local descriptors of protein structure

P. Björkholm[1], F. Lysholm[2], A. Kryshtafovych[3], K. Fidelis[3], and T.R. Hvidsten[4]
[1] – *Stockholm Bioinformatics Center, Stockholm University, Stockholm, Sweden,*
[2] - *IFM Bioinformatics and Swedish e-Science Research Centre (SeRC), Linköping University, Karolinska Institutet, Department of Cell and Molecular Biology, Linköping & Stockholm, Sweden,*
[3] - *UC Davis Genome Centre, UC Davis, USA,*
[4] - *Umeå Plant Science Centre, Department of Plant Physiology, Umeå University, Umeå, Sweden*
torgeir.hvidsten@plantphys.umu.se

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

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

HMMs are used to model local descriptor groups. Each position in the multiple alignment of structurally matching descriptors is modeled as a match state while the rest of the sequence (not matching the local descriptor) is modeled by insert states. Some groups may contain fragments of varying length because only parts of the fragments structurally match the group according to the defined similarity threshold[3]. This is handled by using emitting delete states that are tied to specific match states. In order to ensure that whole fragments are not deleted there are two different types of delete states that are disconnected; delete states that are located in the beginning of the fragments and delete states that are located at the end of the fragments. Our HMMs contain two layers of hidden states; one layer modeling the amino acid content of the local descriptor groups and one modeling the predicted secondary structure. A modified Viterbi algorithm is used to obtain the most probable alignments between a local descriptor group and a target sequences represented by a multiple alignment of related sequences and the predicted secondary structure. We found that the best approach to discriminate targets that contain a local descriptor from targets that do not was to consider the sum of the log values from the match and delete state emissions/transitions only. This solves the problem of comparing the scores obtained from targets with different sequence lengths. Each HMM was matched to predicted domains[5] in the target and accepted if the Viterbi score was higher than an associated threshold shown to discriminate relevant targets in the training set. Contacts between residues located in different backbone fragments were then transferred from the accepted local descriptor groups to the target. Each predicted contact was given a score calculated based on a combination of the scores from all HMMs predicting that contact and the popularity

of that contact in the corresponding local descriptor groups. Thus, contacts predicted by many different local descriptor groups were given a higher weight than contacts predicted by fewer models.

1. Floudas,C.A., Fung,H.K., McAllistera,S.R., Mönnigmanna,M. & Rajgariaa,R. (2006) Advances in protein structure prediction and de novo protein design. *Chemical Engineering Science* **61**, 966-988.
2. Sitao,W., & Yang,Z. (2008) A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. *Bioinformatics* **24**, 924-931.
3. Hvidsten,T. R., Kryshtafovych,A. & Fidelis,K. (2008) Local Descriptors of protein Structure: A systematical analysis of the sequence-structure relationship in proteins using short- and long-range interactions. *Proteins: Structure, Function, and Bioinformatics* **75**, 870-884.
4. Björkholm,P., Daniluk,P., Kryshtafovych,A., Fidelis,K., Andersson,R. & Hvidsten,T.R., (2009) Using multi-data hidden Markov models trained on local neighborhoods of protein structure to predict residue–residue contact. *Bioinformatics* **25**, 1264-1270.
5. Cheng,J. (2007) DOMAC: an accurate, hybrid protein domain prediction server, *Nucleic Acids Res.* **35**, W354-356.

GeneSilico

The GeneSilico pipeline for protein structure prediction

M.J. Boniecki1, E. Wywial1, I. Korneta1, A. Lukasik1, K. Rooijers1, W. Potrzebowski1, M.A. Mika1, M. Korycinski1, K.H. Kaminska1, Ł.P. Kozłowski1, M.J. Pietal1, M. Pawlowski1, J.M. Bujnicki1,2

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

Our pipeline was based around the GeneSilico fold-recognition MetaServer[1] as well as meta-MQAP and REFINER[2] programs. The MetaServer[1] provided a collated view of each target and its structural features (i.e. secondary structures, transmembrane helices, domain architecture, disorder, solvent accessibility, etc). In general, we used templates with high target-template sequence identity and/or predictions from CASP9 servers as starting points for the modeling process.

Methods

The homology-built models and/or predictions from CASP9 servers were superimposed using SWISS-MODEL[3]. Then new models were constructed from the resulting structural alignments using MODELLER[4] and/or SWISS-MODEL[3]. This process was iterated with occasional exclusion or inclusion of models until one or more sufficiently qualitative model(s) were obtained. Next, poorly modeled regions were remodeled using de novo methods (i.e. REFINER[2] and ROSETTA[5]) and loop refinement was performed using REFINER[2]. All decoys were scored using MetaMQAPII[6] and MQAPmulti (unpublished results, see CASP9 abstract). At the end, the best scoring decoys were used to construct hybrid models subjected to global refinement using REFINER[2].

The selection of the five best representative models for a given target was largely dependent on the availability of homology templates for that target. If homology templates were available; then selected models were usually the closest to the templates. If no homology templates were available, five divergent structures were chosen, usually with the aid of MetaMQAPII[6], visual inspection and analysis of possible biological features and functions.

Structural fragment(s) of models for refinement were selected either as per the organizers suggestions or using MetaMQAPII[6]. In all cases, refinement was performed using REFINER[2] and the best models were selected using REFINER[2] as well as MetaMQAPII[6]. Occasionally models were improved using MODELLER[4], and then re-scored.

1. Kurowski MA, Bujnicki JM. (2003) GeneSilico protein structure prediction meta-server. *Nucleic Acids Res.* **31**(13):3305-3307.
2. Boniecki M, Rotkiewicz P, Skolnick J, Kolinski A. (2003) Protein fragment reconstruction using various modeling techniques. *J Comput Aided Mol Des.* **17**(11):725-738.
3. Arnold K, Bordoli L, Kopp J, Schwede T. (2006) The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* (Oxford, England) **22**(2):195-201.
4. Sali A, Blundell TL. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**(3):779-815.
5. Das R, Baker D. (2008) Macromolecular modeling with rosetta. *Annu Rev Biochem.* **77**:363-382.
6. Pawlowski M, Gajda MJ, Matlak R, Bujnicki JM. (2008) MetaMQAP: a meta-server for the quality assessment of protein models. *BMC bioinformatics* **9**:403.

GOBA_Wroc_PL
GOBA_PL_07

Assessment of model quality based on protein structural and functional similarities

B.M. Konopka, W. Dyrka, M. Rybicka, P. Gasior[1], J.-C. Nebel[2] and M. Kotulska[1]
Institute of Biomedical Engineering & Instrumentation, Wroclaw University of Technology, Poland
Faculty of Computing, Information Systems & Mathematics, Kingston University, United Kingdom
malgorzata.kotulska@pwr.wroc.pl

GOBA (Gene Ontology-Based quality Assessment) is a Model Quality Assessment Program that evaluates the quality of a predicted protein structure by comparing the function of the target protein with that of its structural neighbors. Predictions in QA and TS human categories were submitted.

Methods

In the proposed approach, the evaluation of the quality of a predicted protein structure is achieved by quantifying its functional similarity with those of structural neighbors. This scheme relies on the standardized description of protein functions provided by the Gene Ontology (GO)[1], which allows quantitative measures of functional similarity. Consequently, this method can only be applied to target proteins associated with GO term annotations.

The model quality assessment procedure is the following. First, the predicted structure is used as an input of DALI_lite[2] which conducts a search for structural neighbors (SNs) in the DALI database of protein structures (http://ekhidna.biocenter.helsinki.fi/dali_lite/downloads/v3/). This returns a list of structures associated with Z-scores. Then, the semantic similarity between the protein functions of SNs and the target protein is calculated using Wang's algorithm[3]. According to Wang scores, structural neighbors are divided into negative and positive sets. Finally, a Receiver Operating Characteristic[4] (ROC) curve is plotted based on the DALI Z-scores. The area under the ROC curve (AUC) is used to measure the quality of the model: AUCs of best models should approach one, while the worst models should score around 0.5.

The GOBA_PL_07 group tested the basic approach using a functional similarity threshold of 0.7. The GOBA_Wroc_PL group used the average of a set of AUCs generated for thresholds between 0.00 – 1.00 with a step of 0.01.

In the TS category, a number of online servers were chosen to provide candidate structures**.** The models were ranked by the AUC GOBA measures and the TOP5 were submitted to the contest.

Results

Out of the 60 issued targets, 18, respectively 17, were processed by GOBA_Wroc_PL and GOBA_PL_07, (85 and 81 models were submitted in the TS category). Out of the 15 servers that were used to generate the pool of models, four were the most successful in providing models that qualified to the submitted TOP5: AS2TS, Phyre, GeneSilico Metaserver, SAM_T08.

In the QA category 79 prediction servers were evaluated. Although the rankings produced by the two tested methods based on the average scores of the best submitted models differed quite significantly, both methods were fairly consistent in predicting top performing servers (Table 1, Table 2). The best servers according to GOBA were: MULTICOM-CLUSTER, CLEF-Server and FALCON-SWIFT.

Table 1: TOP5 servers according to GOBA_Wroc_PL

Rank	ID	Group name	Average	SD	#of Targets
1	11	CLEF-Server	0.694	0.142	18
2	14	FALCON-SWIFT	0.694	0.142	18
3	38	MULTICOM-CLUSTER	0.686	0.167	17
4	39	MULTICOM-CONSTRUCT	0.684	0.169	17
5	41	MULTICOM-REFINE	0.683	0.164	17

Table 2: TOP5 servers according to GOBA_PL_07

Rank	ID	Group name	Average	SD	#of Targets
1	77	YASARA	0.812	0.103	7
2	38	MULTICOM-CLUSTER	0.811	0.169	16
3	11	CLEF-Server	0.801	0.161	17
4	14	FALCON-SWIFT	0.801	0.161	17
5	29	Jiang_THREADER	0.784	0.176	16

Availability

The source code and linux binaries of the application are freely available upon email request.

1. The Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. Nat. Genet. 25(1), 25-29.
2. Holm L, Rosenström P (2010) Dali server: conservation mapping in 3D. Nucl. Acids Res. 38, W545-549
3. Wang J. Z., Du Z., Payattakool R., Yu P. S. and Chen C-F (2007) A new method to measure the semantic similarity of GO terms. Bioinformatics 23 (10), 1274–1281
4. C.E. Metz (1978) Basic principles of ROC analysis. Semin Nucl Med. 8(4), 283-98

GSmetaDisorder
GSmetaDisorder3D
GSmetaDisorderMD
GSmetaserver

Meta-prediction of intrinsic disorder in proteins using different sources of information

Ł.P. Kozłowski[1] and J.M. Bujnicki[1,2]

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

Protein disorder prediction is an important step during elucidating protein function and in the last years became standard procedure before protein structure determination. To date, more than 30 disorder prediction methods have been created among the many types of bioinformatics tools designed to analyze proteins features. Encouraged by the success of our meta prediction method (GSmetaDisorder) in CASP8, we extended the approach to more diverse sources of information. We constructed a set of methods which validate each other while they incorporate fold recognition (FR) programs.

Methods

GSmetaDisorder - this method was used for the first time in CASP8 (three different variants tested, first two places in disorder prediction in CASP8[1]). It uses 13 primary disorder predictors (DisEMBL[2], DISOPRED2[3], DISpro[4], Globplot[5], iPDA[6], IUPred[7], Pdisorder (Softberry, Inc.), Poodle-s[8], Poodle-l[9], PrDOS[10], Spritz[11], DisPSSMP[12,] and RONN[13]) to construct the final consensus result (for more details see the CASP8 abstracts book).

GSmetaDisorder3D – a naive predictor, which tries to deduce the presence of disorder by counting gaps in alignments produced by fold recognition methods (HHSEARCH[14] run over PDB70 and CDD databases, FFAS[15], MgenThreader[16], PSI-BLAST[17] run in two different modes on the top of cullpdb, PHYRE[18] and PCONS5[19] which uses models from previous methods as an input). To address the problem of accuracy of FR methods and different reliability of hits, we used a genetic algorithm implemented in Pyevolve[20]. This procedure produced weights for each FR method depending whether hits are reliable, less reliable or below statistical importance. Obtained weights have been used to construct a consensus method. The method was trained on CASP8 targets.

GSmetaDisorderMD - uses a genetic algorithm to combine GSmetaDisorder with GSmetaDisorder3D.

GSmetaserver - the same as GSmetaDisorderMD, but uses a different score for the optimisation of weights by the genetic algorithm. The score called Sww is a mixture of a standard Sw score used by CASP assessors and the classical AUC. In principle, it is calculated like the AUC, but with TPR and FPR are substituted by Sw score.

Results

GSmetaDisorder was tested again to compare it with the three new methods (AUC > 0.91 when tested during CASP8). GSmetaDisorder3D was constructed with the aim of calculating how much information about protein disorder can be extracted from the output of FR methods alone (AUC = 0.87). However, the main question addressed with GSmetaDisorderMD was whether the information from FR predictions can help to improve the result based on the primary disorder predictors. According to the test done on the CASP8 dataset, GSmetaDisorderMD achieves the AUC of 0.93, which clearly shows that adding information about template coverage detected by FR methods contributes to the successful prediction. Finally, the difference between the performance of GSmetaserverMD and GSmetaserver was statistically insignificant. This result shows that the robustness of the method depends mostly on the information extracted from the output of primary predictors (both disorder predictors and FR methods) and it is not dependent on the optimization procedure.

Availability

The meta prediction server can be accessed from http://iimcb.genesilico.pl/metadisorder/

1. Noivirt-Brik,O., Prilusky,J. & Sussman,J.L. (2009). Assessment of disorder predictions in CASP8. Proteins 77 Suppl 9, 210-6.
2. Linding,R., Jensen,L.J., Diella,F., Bork,P., Gibson,T.J. & Russell,R.B. (2003). Protein disorder prediction: implications for structural proteomics. Structure 11, 1453-9.
3. 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, 2138-9.
4. Hecker,J., Yang,J.Y. & Cheng,J. (2008). Protein disorder prediction at multiple levels of sensitivity and specificity. BMC Genomics 9 Suppl 1, S9.
5. Linding,R., Russell,R.B., Neduva,V. & Gibson,T.J. (2003). GlobPlot: Exploring protein sequences for globularity and disorder. Nucleic Acids Res 31, 3701-8.
6. Su,C.T., Chen,C.Y. & Hsu,C.M. (2007). iPDA: integrated protein disorder analyzer. Nucleic Acids Res 35, W465-72.
7. Dosztanyi,Z., Csizmok,V., Tompa,P. & Simon,I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 21, 3433-4.
8. Shimizu,K., Hirose,S. & Noguchi,T. (2007). POODLE-S: web application for predicting protein disorder by using physicochemical features and reduced amino acid set of a position-specific scoring matrix. Bioinformatics 23, 2337-8.
9. Hirose,S., Shimizu,K., Kanai,S., Kuroda,Y. & Noguchi,T. (2007). POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions. Bioinformatics 23, 2046-53.
10. Ishida,T. & Kinoshita,K. (2007). PrDOS: prediction of disordered protein regions from amino acid sequence. Nucleic Acids Res 35, W460-4.
11. Vullo,A., Bortolami,O., Pollastri,G. & Tosatto,S.C. (2006). Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines. Nucleic Acids Res 34, W164-8.
12. 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.
13. Yang,Z.R., Thomson,R., McNeil,P. & Esnouf,R.M. (2005). RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. Bioinformatics 21, 3369-76.
14. Soding,J. (2005). Protein homology detection by HMM-HMM comparison. Bioinformatics 21, 951-60.
15. Jaroszewski,L., Rychlewski,L., Li,Z., Li,W. & Godzik,A. (2005). FFAS03: a server for profile--profile sequence alignments. Nucleic Acids Res 33, W284-8.

16. 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.
17. 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.
18. Bennett-Lovsey,R.M., Herbert,A.D., Sternberg,M.J. & Kelley,L.A. (2007). Exploring the extremes of sequence/structure space with ensemble fold recognition in the program Phyre. Proteins.
19. Wallner,B. & Elofsson,A. (2005). Pcons5: combining consensus, structural evaluation and fold recognition scores. Bioinformatics 21, 4248-54.
20. Butterfield,A., Vedagiri,V., Lang,E., Lawrence,C., Wakefield,M.J., Isaev,A. & Huttley,G.A. (2004). PyEvolve: a toolkit for statistical modelling of molecular evolution. BMC Bioinformatics 5, 1.

Hamilton_Huber

Protein contact prediction using patterns of correlation

N. Hamilton[1] and T.Huber[2]
[1] -*Institute for Molecular Bioscience, The University of Queensland, St. Lucia, Queensland, Australia,* [2] – *Research School of Chemistry, Australian National University, Canberra, ACT, Australia*
n.hamilton@imb.uq.edu.au, t.huber@anu.edu.au

Protein contact prediction provides a complementary approach to the information provided by force field and sequence alignment based methods for protein fold prediction. While the predictive accuracy is far from perfect it can provide valuable information that can be used, for instance, to rank models created by other methods. In the following we describe PoCM, a new method for contact prediction by training a Neural Network to classify patterns of contact[1]. The main inputs to the neural network are a set of 25 measures of correlated mutation between all pairs of residues in two "windows" centered on the residues of interest. The individual pairwise correlations are a relatively weak predictor of contact, but by training the network on windows of correlation the accuracy of prediction is significantly improved.

Methods

The Psipred[2] version 2.3 software is used to generate a prediction for the secondary structure as well as giving a pair-wise multiple sequence alignment for the proteins sequence. For each pair of residues in the protein sequence we generate a pattern of inputs for a neural network as follows.

Pairwise correlations. The multiple sequence alignment is used to calculate the (mutational) correlation between two columns of the multiple sequence alignment. The correlations are calculated as in Göbel et al.[3], with the minor modification that the Blosum62 matrix rather than that of McLachlan is used to score the residue interchanges. Windows of length 5 of consecutive columns are found. For each pair of non-overlapping windows the 25 correlations between columns of the first window with columns of the second are used as inputs to the neural network. The aim is to predict whether the middle residue of the first window is in contact with the middle residue of the second.

Residue classes. Residues may be classified as non-polar, polar, acidic, or basic. For a pair of residues there are ten possible pair cases. Thus we have ten binary inputs, exactly one of which is set to one to encode the residue type of the pair we are attempting to predict on.

Predicted secondary structure. For a given residue, its predicted secondary structure type is encoded as three binary inputs, being either helix, sheet or neither. For a given residue pair that we are attempting to predict with, the predicted secondary structure is input for the two residues as well as the two residues that are adjacent to them.

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

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

Network Architecture and Training

The predictor neural network is a standard feed-forward network, with 56 inputs, ten hidden units, and a single output. The expected output is 1 for contacts and 0 for non-contacts. The network was trained, validated and tested on disjoint sets of 100, 50 and 1033 proteins using back propagation with a momentum term with the Stuttgart Neural Network Simulator[4].

Results

The trained network was tested on a set of 1033 proteins of known structure. An average predictive accuracy of 21.7% was obtained taking the best L/2 predictions for each protein, where L is the sequence length. Taking the best L/10 predictions gives an average accuracy of 30.7%. Similar accuracies were obtained in independent blind tests of CASP7 and CASP8.

Availability

The automated prediction server can be found at
http://newcompbio.biosci.uq.edu.au/~huber/PoCM/contact_casp9.html

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

Handl_Lovell
Lovell_group

Iterated local search approaches to de novo prediction with Rosetta

J. Handl and S.C. Lovell
University of Manchester
j.handl@manchester.ac.uk

We describe two different variants of an iterated local search scheme that have been designed as alternatives to the traditional Monte Carlo search in Rosetta's low resolution sampling protocol. The design of these techniques has been motivated by our belief that the development of more refined methods for the exploration and traversal of energy landscapes is essential for further progress in ab initio protein structure prediction, and that this holds true both at the low-resolution and the all-atom energy stage.

Methods

In the previous CASP experiment, we described an iterated local search (ILS) approach that wrapped around the established prediction method, Rosetta[1] using it as a "local" search routine. The aim was to attempt to "learn" from the results of individual runs of Rosetta and to improve upon the performance of random restarts. Specifically, candidate structures generated by individual runs of Rosetta were evaluated with regard to four objective functions (Rosetta's low resolution energy, its short-range and long-range hydrogen term and its radius of gyration), and Rosetta was restarted from structures that performed (Pareto) optimally with respect to these terms. The resulting method was successful at generating decoys with energies significantly lower than those obtained by random restarts of Rosetta. These lower energies did not consistently translate into lower-RMSD structures, which was caused by two factors: the existence of deep local optima in the low-resolution Rosetta energy function and inadequate sampling of conformational space by individual runs of Rosetta in our ILS protocol[2].

In this CASP, we have experimented with additional mechanisms to increase the diversity of conformational sampling and facilitate the escape from local optima. Rather than wrapping around Rosetta, we directly modified and integrated our new iterated local search protocols into Rosetta's low-resolution stage. In particular, two different variants were proposed.

Lovell_group. The primary technique tested in this CASP, integrates an "archive" of good solutions with Rosetta's Monte Carlo sampling technique. For this purpose, the standard Rosetta low-resolution protocol was modified in a number of ways: (i) Instead of a sequence of scoring functions, a single scoring function (using nine out of ten of Rosetta's low resolution energy terms) was used. (ii) The temperature of the Monte Carlo search was kept constant and stagnation of the search was now used as an indication to compare the current solution to the archive and restart the search. (iii) An archive of structures that are non-dominated with respect to a set of three groups of objectives was maintained. New starting points for the search were obtained from this archive through the use of a crossover or destruction operator. The crossover operator combined features from two structures, whilst the destruction operator perturbed a designated part of a given conformation. At the end of the search. the content of the archive was returned rather than a single solution.

Hand_Lovell. The second technique tested in this CASP also implemented the above changes, but introduced a fifth modification. A multiobjective hillclimber replaced the single-objective Monte Carlo search as the local search technique. In the implementation tested during CASP, this multiobjective

hillclimber used the same set of objectives as employed in the archive, but an independent choice would also be feasible.

Results

During the experiment, the above two techniques were used to generate predictions for all targets in the human/server categories in the same fashion. For each target, the two algorithms were run 10 times, where the runtime of one run of the algorithms corresponds roughly to about 100 standard runs of the low-resolution stage of Rosetta. The final archives from all ten runs were combined and our Model 1 to Model 4 submissions were selected from this set in an automated way (based on their performance with respect to specific low-resolution energy terms). Model 5 was used as a control.

For the first method (Lovell_group), the energy values obtained appeared to indicate that the exploration of the search space may have improved compared to standard Rosetta runs. For the second method (Handl_Lovell), the energy values indicated a performance decrease, particularly so for larger structures.

Availability

1. Simons, K.T., Kooperberg C., Huang, E., Baker, D. (1997) Assembly of protein tertiary structures from fragments with similar local sequences. *J. Mol. Biol.*. **268**, 209-25
2. Handl, J. and Lovell, S. (2008) De novo prediction using multiobjective iterated local search. CASP abstracts. 49-50

HEU_DisIP

Predicting intrinsically disordered regions based on a ensemble method

Bo He, Weixing Feng, Yan Fan, Kejun Wang
wangkejun@hrbeu.edu.cn

In this round of CASP experiment (CASP9), we used a method of combining multiple predictors to identify Intrinsically Disordered Regions (IDPs) from amino acid sequences. In the design process of our predictors, we made use of the idea of an ensemble method. It could combine not only the predicting results from several predictors of IDRs based on every computational technique but also the predicting results from different computational techniques.

In our predictor, a total of 420 features could be generated from amino acid sequences to predict IDRs. The first 20 features were derived from the statistics of 20 kinds of amino acids within a given sliding window. Then we introduced next 400 features from the statistics of dimer amino acids representing the number of the pattern of each amino acid followed by another amino acid in the window.

At present, in the predicting field of IDPs, the most common computational techniques are ANN and SVM[1]. So we selected SVM and two kinds of neural network including BP and RBF to build the predictors of IDRs based on the above features respectively. Because the performance based on combining several neural networks is better than a single individual according to LK. Hansen and P. Salamon[2], every computational technique was used to make five times tests and build five subpredictors of IDRs in order to abtain the better predicting performance. In these five tests, the used computational technique was performed by setting five different parameters of itself. Thus, we could achieve 15 groups of predicting results for three computational techniques.

At last, the final result was made decision fusion by voting method. First we needed to count the results of disordered structure and the results of ordered structure respectively. If the number of the disordered structure was more than the number of the ordered structure, the final predicting results was to tend to disordered structure. The predicting value was the average of disordered predicting values from subpredictors with disordered results. If the number of the disordered structure was less than the number of the ordered structure, the final predicting results was to tend to ordered structure. The predicting value was the average of ordered predicting values from subpredictors with ordered results.

1. He, B., Wang, K., Liu, Y., Xue, B., Uversky, V. N. & Dunker, A. K. (2009). Predicting intrinsic disorder in proteins: an overview. Cell Res 19, 929-49.
2. Hansen, L. & Salamon, P. (1990). Neural network ensemble; IEEE Trans Pattern Anal Machine Intel 12, 993-1001.

HHpred

Homology based structure prediction by HMM-HMM comparison

A. Meier, M. Remmert, C. Angermueller and J.Soeding
Gene Center, Ludwig-Maximilians University, Munich
[meier, remmert, angermueller, soeding]@genzentrum.lmu.de

For CASP9 we strove to improve on model quality while further shortening response times to below 10 min. In CASP8, our servers were impaired by a bug in the treatment of multiple domains: When evaluated on single-domain targets, HHpred5 was the 2nd best server (according to the official GDT_HA Z-scores), whereas on all targets it occupied rank 6[1]. We sought to overcome this problem as described in point 3. below. The following changes were implemented:

1. To build alignments for query and database sequences, we now use our new iterative HMM-HMM sequence search program HHblits instead of our Perl script buildali.pl, which ran up to 8 iterations of PSI-BLAST. HHblits is faster than PSI-BLAST while being significantly more sensitive and producing multiple alignments of higher quality in terms of sensitivity and precision (Remmert M, Biegert A, Hauser A, Soeding J, unpublished). The query alignment is converted into an HMM with hhmake, and hhsearch from the HHsearch package[2] is used to search for templates in representative HMMs of the PDB (70% maximum sequence identity).

2. Starting with a list of possible templates generated by HHsearch, the final templates are selected with a heuristic approach which tries to produce the maximum coverage of the query with a limited number of templates. We measure coverage of query residues quantitatively in term of the posterior probability for the query residue to be correctly aligned to the template, as calculated from the maximum accuracy alignment algorithm implemented in HHsearch.

3. We replace MODELLER's[3] distance restraints to account for the varying confidence of aligned residue pairs along the alignment, again measured by the posterior probabilities. We define a new type of distance restraint in MODELLER as a mixture of two Gaussians, the two components describing correctly and incorrectly aligned residues. The mixture parameters (means, standard deviations and mixture weights) are predicted by a mixture density network[4], a neural network designed for training the parameters of a mixture of Gaussians. Badly aligned residues with low posteriors will lead to mixtures with flatter components with an increased background mixture weight. Distance restraints from multiple templates are combined by multiplication of probability densities and not by addition as in MODELLER, leading to an reinforcement of correct restraints and a weakening of conflicting restraints.

HHpredA, HHpredB and HHpredC were intended to run extensions to our basic pipeline, which we did not get implemented in time. Therefore, all three servers should have performed identical calculations. A bug in the multi-threading part of our new program HHblits led to bad query alignments in rare cases, which resulted in a few suboptimal models in particular for HHpredC.

Our functional site prediction server HHfuncs ran under the name of HHpredA. It searches for homologous templates in the FireDB database[5] using HHsearch. For each annotated site in the matched FireDB templates, HHfuncs calculates a probability that the site annotation can be transferred, by multiplying three probabilities: (1) HHsearch's probability for a homologous match, (2) the probability that the binding site residues are correctly annotated in FireDB, derived from analyzing their FRpred

scores[6], and (3) the probability that the binding of the functional site is evolutionarily conserved between target and template. If no binding site can be identified with probability of >35%, FRpred is run. If no reliable model could be built by HHpredA (DOPE score < 0.5), the 5 top-ranked residues are predicted to form a functional site. If a 3D model for the protein could be predicted with HHpredA, we employ the RankProp algorithm (Weston J et al. PNAS 2004) to identify spatial clusters of high scores and predict the top-ranking 5 residues as functional site.

1. Hildebrand A., Remmert M., Biegert A., and Soding J. (2009): Fast and accurate automatic structure prediction with HHpred. Proteins. 77 Suppl 9:128-132.
2. Soding J. (2003): Protein homology detection by HMM-HMM comparison. Bioinformatics 21:951-960.
3. Sali A., Blundell T.L. (1993): Comparative protein modelling by satisfaction of spatial restraints. J Mol. Biol. 234:779-815.
4. Bishop C.M. (1994): Mixture Density Networks. Neural Computing Research Group Report NCRG/94/004.
5. Lopez G., Valencia A., Tress M. (2007): FireDB -a database of functionally important residues from proteins of known structure. Nucl. Acids Res., Vol. 35, No. suppl_1., pp. D219-223
6. Fischer J., Mayer C. and Soding J. (2008): Prediction of protein functional residues from sequence by probability density estimation. Bioinformatics 24:613-620.

HIT_Dict

Prediction of intrinsically disordered regions with statistical dictionaries

Wei Yang, Kuan-Quan Wang and Wang-Meng Zuo
Biocomputing Research Centre, School of Computer Science and Technology, Harbin Institute of Technology, Harbin, 150001, China
wangkq@hit.edu.cn

We present a novel method, Diso_Dict, to use both homologous and non-homologous information for protein disorder prediction. Using the known protein disorder information available in the PDB database[1; 2], Diso_Dict first constructs a set of statistical dictionaries which contain the frequency information of disorder and order with a well-defined linked list structure, then combines the statistical dictionaries with the PSSM scoring matrix to fast generate the neighbor list, and finally uses the maximum score rule to predict disordered region.

Methods

Construction of statistical dictionaries

Let $R = (R_1, \ldots, R_i, \ldots, R_n)$ be a peptide fragment of length n. A statistical dictionary is used to record the occurrence frequencies of the disorder or order states corresponding to the ith residue (R_i) for all possible R in the dataset of protein with known protein disorder information. Since protein is composed of 20 different amino acids, there are 20^n possible sequences for the peptide fragment of length n. In order to denote any possible peptide fragments with a given length, we first encode 20 amino acids A, R, ..., V with 0, 1, ..., 19, respectively, and thus a peptide fragment can be regarded as a 20-ary number. Then, every peptide fragment can be uniquely represented by a number between 0 and 20^n-1, which is called the encoding of the peptide fragment. For example, the encoding of the peptide fragment ARRVV is 8799. Therefore, when the frequency information of two state types corresponding to each peptide fragment is stored in an array or a file indexed by the encoding of the peptide fragment, it does not require any space to store the information of amino acids in the peptide fragment. However, the statistical dictionary of length n constructed in this way needs 4×20^n B of storage space if two bytes are used to store the frequency of each state type. In order to save the storage space, we use a linked list structure to construct statistical dictionary for the peptide fragment whose length is larger than 5. For fast frequency information retrieval, each peptide fragment of length n is divided into two parts: the five central consecutive amino acid residues (offset peptide) and the other n-5 amino acid residues (identity peptide). For example, the offset peptide and identity peptide of the peptide fragment VARRFFA are ARRFF and VA, respectively. In this way, an array of pointer of length 20^5 is first constructed. Then, the frequency information of the peptide fragment in the dataset of protein with known structure can be quickly inserted into a linked list according to the encoding of the identity peptide in descending order, where the index of the head pointer of the linked list in the array of pointer equals the encoding of the offset peptide.

Generation of similar list

It is known that the size of the known protein sequence data is significantly larger than that of the protein data with known disorder state. Thus, for many peptide fragments, there is no corresponding structure information in the protein structure database. Therefore, when the protein chain to be predicted contains such peptide fragments, the statistical dictionary fails to give effective frequency information.

Considering that the structure of protein is more conservative than its sequence, it would be beneficial to reference the structure information of the peptide fragments which are similar to them.

Appropriate definition of similarity is critical in the generation of similar list. The two peptide fragments *p* and *q* are regarded as similar, if they have the same length and the scores corresponding to all the residue pairs at the same position are positive in terms of the specific scoring matrix. The similar score of two peptide fragment *p* and *q* is defined as the sum of the scores of all the amino acid pairs at the same position. The similar list of *p* is composed of all the peptide fragments similar to it. Specifically, the PSSM scoring matrix is used in our work.

According to the above definitions, it is easy to generate the similar list of a given peptide fragment. First, PSSM scoring matrix is preprocessed with the following three steps: (1) combine each score with the amino acid which lies in the same column with it as a whole; (2) delete all the combinations with non-positive score; (3) sort the remaining combinations of each row according to score in descending order. After preprocessing, the row of the scoring matrix lists all feasible substitutions of its corresponding amino acid. Second, amino acid positioning is used to find the row corresponding to each amino acid in the given peptide fragment. Finally, the similar list is constructed by traversing all feasible substitutions of amino acids in the given peptide fragment.

The Diso_Dict Algorithm

Based on the constructed statistical dictionaries, the Diso_Dict algorithm is developed to predict the protein disordered regions. Diso_Dict associates each residue of the query protein with two confidence scores: *CScore*(*D*) and *CScore*(*O*), which correspond to the disorder and order state types, respectively. The calculation of the confidence score could be divided into two cases:

(1) For the internal residues of the query protein chain, a sliding window of length n (n>5) is used to scan it. If the peptide fragment in the sliding window could be found in the statistical dictionary, the normalized frequencies are directly assigned to the residue as its confidence scores. Otherwise, a neighbor list is constructed to compute the confidence scores.
(2) For the remaining residues, the calculation of the confidence scores is similar to that of the internal residue except the choice of window length and statistical dictionary.

After obtaining the confidence scores, the maximum score rule, which assigns the state type with the maximum confidence score to the target residue, is used to predict disordered regions.

1. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res* **28**, 235-42.
2. http://www.rcsb.org/pdb/files/ss_dis.txt.gz.

Infobiotics

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

J. Bacardit[1,2], P.Widera[1] and N. Krasnogor[1]

1 - School of Computer Science, University of Nottingham,
2 – School of Biosciences, University of Nottingham
jaume.bacardit@nottingham.ac.uk

Our method is identical to the one that participated in the previous CASP edition, except for the usage of a much larger training set. We have constructed an ensemble of more than one thousand rule sets to participate in the residue-residue contact category of CASP. The rule sets were generated by BioHEL[1], our in-house machine learning system. Three types of input information were used to train our system: (1) detailed local sequence information from three selected regions (windows) around specific residues, (2) information about the connecting segment between the two target residues and (3) global sequence information.

There were two windows of ±4 residues around the two target residues and a window of ±2 around the middle point in the chain between the two target residues4 Residue in the three windows was characterised characterised using (1) a position-specific scoring matrix (PSSM) profile computed with PSI-BLAST[3], (2) secondary structure predicted by PSIPRED[5], (3) five-state coordination number (CN)[6], (4) five-state relative solvent accessibility (SA)[1]and (5) five-state Recursive Convex Hull (RCH)1, all three predicted using BioHEL.

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

The training process followed the four steps below:

1. We selected a set of 3262 (2811 in CASP8) protein chains from PDB-REPRDB with a resolution less than 2Å, less than 30% sequence identity and without chain breaks nor non-standard residues. We used 90% of the proteins (~573000 residues) for training and 10% for test. This training set was used to predict RCH, SA and CN.
2. For the residue-residue contact prediction, the size of the training set was reduced: All proteins with less than 250 residues and only a random 20% of proteins longer than 250 residues were kept. Still, the new set contained 32 million pairs of residues (15.2M in CASP8), from which less than 2% were real contacts.
3. To balance the training set (in terms of contacts/non contacts) we created 50 random samples from these 32 million pairs. Each sample contained around 660000 residue pairs (300000 in CASP8) with a fixed 2:1 proportion of non-contacts to real contacts.

4. We run BioHEL 25 times for each training sample with different initial random seeds, thus generating an ensemble of 1250 rule sets (50 training samples x 25 seeds) to perform the residue-residue contact prediction.

1. M. Stout, J. Bacardit, J.D. Hirst and N. Krasnogor. (2008) Prediction of Recursive Convex Hull Assignments for Protein Residues. Bioinformatics 24(7):916-923.
2. T. Noguchi, H. Matsuda, and Y. Akiyama. (2001). Pdb-reprdb: a database of representative protein chains from the protein data bank (pdb). *Nucleic Acids Res*, 29:219–220.
3. S.F. Altschul, , T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, & D.J. Lipman, (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
4. M. Punta and B. Rost (2005) • gProfcon: novel prediction of long-range contacts • h. Bioinformatics 21(13):2960-8.
5. D.T. Jones, (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol. **292***, 195-202.
6. J Bacardit, M. Stout, J.D. Hirst, N. Krasnogor and J. Blazewicz. (2006) Coordination Number Prediction using Learning Classifier Systems: Performance and Interpretability. Proceedings of the 8th Annual Conference on Genetic and Evolutionary Computation (GECCO2006), pp. 247-254, ACM Press
7. G. Shackelford and K. Karplus. (2007) Contact Prediction using Mutual Information and Neural Nets. *Proteins: Structure, Function, and Bioinformatics*, 69(S8):159-164.

IntFOLD-TS
IntFOLD-DR
IntFOLD-FN
IntFOLD-QA

Fully Automated Prediction of Tertiary Structure, Disorder, Binding Site Residues and Model Quality Using the IntFOLD Server

D.B. Roche[1] and L.J. McGuffin[1]
[1] - *School of Biological Sciences, University of Reading, Whiteknights, Reading RG6 6AS, UK*
l.j.mcguffin@reading.ac.uk

The IntFOLD server was newly developed for CASP9 and integrates the latest versions of our new automated methods for fold recognition (nFOLD4), disorder prediction (DISOclust 2.0), binding site residue predictions (FunFOLD) and model quality assessment (ModFOLD 3.0).

Methods

For CASP9, a bespoke version of the server was developed in order to return results for each category (TS, DR, FN, QA), hence 4 alternative groups were registered for the IntFOLD server.

IntFOLD-TS: nFOLD4

In order to generate TS predictions the IntFOLD server implemented the latest version of the nFOLD method[1]. The nFOLD4 method works by integrating the alignment output from the SP3[3], SPARKS[4], HHsearch[5] and COMA[6] methods and then generating around 40 alternative multiple and single template based 3D models using Modeller[7]. For each target, all the generated models were then ranked using the ModFOLDclust2 QA method[2] and the top 5 were submitted.

IntFOLD-DR: DISOclust 2.0

The latest version of our DISOclust[8] method was used to generate automated DR submissions via the IntFOLD server. The new method uses the ModFOLDclust2 QMODE2 output in order to identify the regions of high variability occurring in nFOLD4 models.

IntFOLD-FN: FunFOLD

The FunFOLD method uses structural superpositions of the top ranked nFOLD4 3D models and related templates with bound ligands in order to identify putative contacting residues. The methods uses a novel fully automated approach for both ligand cluster identification and residue selection and it is competitive with the best manual FN methods that were tested at CASP8.

IntFOLD-QA: ModFOLD 3.0

Finally, the IntFOLD server also integrates the ModFOLD 3.0 QA method. This new version of ModFOLD is capable of carrying out either single-model mode or multiple-model mode clustering. Each model is compared against the models generated by nFOLD4 (and any other provided models) using the ModFOLDclust2 method.

Availability

An alpha version of the IntFOLD server with graphical output is available at: http://www.reading.ac.uk/bioinf/IntFOLD/IntFOLD_form.html.

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), 143-151.
2. McGuffin,L.J. & Roche,D.B. (2010) Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments. *Bioinformatics*. **26**, 182-188.
3. McGuffin,L.J. (2008) Intrinsic disorder prediction from the analysis of multiple protein fold recognition models. *Bioinformatics*. **24**, 1798-1804.
4. Zhou,H. & Zhou,Y. (2005) SPARKS 2 and SP3 servers in CASP6. *Proteins*. **61** (S7), 152-156.
5. Söding, J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*. **21**, 951-96.
6. Margelevičius,M. & Venclovas Č. (2010) Detection of distant evolutionary relationships between protein families using theory of sequence profile-profile comparisons. *BMC Bioinformatics*. **11**, 89.
7. Sali,A. & Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol*. **234**, 779-815.
8. McGuffin, L.J. (2008) Intrinsic disorder prediction from the analysis of multiple protein fold recognition models. *Bioinformatics*. **24**, 1798-1804.

IsUnstruct

IsUnstruct: a method based on a model inspired by the Ising model for prediction of disordered residues from protein sequence alone

M.Yu. Lobanov, O.V. Galzitskaya
Institute of Protein Research of the Russian Academy of Sciences, 4 Institutskaya str., Pushchino, Moscow Region, 142290, Russia
ogalzit@vega.protres.ru

Intrinsically disordered regions serve as molecular recognition elements, which play an important role in the control of many cellular processes and signaling pathways. It is useful to be able to predict positions of disordered residues and disordered regions in protein chains using protein sequence alone.

Methods

The presented method (named IsUnstruct) allows the prediction of disordered regions and disordered residues in a protein molecule starting solely from its sequence. It is based on a simple physical model. According to this model, each residue can be in one of two states: ordered (structured or fixed) or disordered (unstructured or free). Experimentally, fixed residues are determined by X-ray analysis, but free ones are not resolved. The model is an approximation of the Ising model[1] in which the interaction term between neighbors has been replaced by a penalty for the state change (the energy of border). This allows us to apply dynamic programming to the Ising problem. Our systematic analysis of disordered regions in PDB (2008) revealed 345 disordered patterns of different length[2]. We took into account this library of disordered patters.

Results

The IsUnstruct has been compared with other available methods and found to perform well. The method correctly finds 77% of disordered residues as well as 87% of ordered residues in the CASP8 database, and 72% of disordered residues as well as 85% of ordered residues in the Disprot database (version 5.00, which includes 517 protein chains).

Availability

http://antares.protres.ru/IsUnstruct/

1. Ising,E. (1925) Beitrag zut Theorie des Ferromagnetizmus. *Zeitschr Phys **31***, 253-258.
2. Lobanov,M.Yu., Furletova,E.I., Bogatyreva,N.S., Roytberg,M.A., Galzitskaya,O.V. (2010) Library of Disordered Patterns in 3D Protein Structures. *PLoS Computational Biology* **in press**.

I-TASSER-FUNCTION

See Zhang_FUNCTION

JAMMING

Prediction of interface residues based on network connectivity

G. Del Rio[1] and R. Corral Corral[1]
[1] – *Department of Biochemistry and Structural Biology; Instituto de Fisiología Celular/UNAM 04510 México DF, México*
gdelrio@ifc.unam.mx

Interface residues of protein complexes hold key information on protein function. Thus, prediction of interface residues constitutes a way to evaluate our understanding about the structure-function relationship in proteins and, different features observed at protein interfaces have been used to predict them but residue centrality, a feature known to be related to functional residues in proteins. In this work we present our results predicting interface residues using network centrality.

Methods

Only target sequences with a clear homologue in the PFAM[1] database were considered for this experiment. Multiple sequence alignments (MSA) from PFAM were aligned to the target sequence using MUSCLE[2]. From this alignment, 10 different atomic models for each target protein sequence were generated using MODELLER9v8[3]. No further refinement was performed on these models. For each target sequence, a list of conserved residues was derived from its MSA. For each atomic model, 500 conformers were derived using ElNemo[4] and their central residues were determined using JAMMING[5]. A score for each of the ten atomic models was obtained by comparing the probability to identify a conserved residue from the central residues. The best-scored model was further used to predict interface residues.

All the central residues for a given atomic model were grouped in a single set, connected subgraphs or cliques. Proteins reported in the MolMov[6] database and the SCOP[7] database were used as training sets. Different clustering algorithms implemented in the WeKa[8] java library were used to identify interface residues in the SCOP training set. To validate the prediction in the training sets, residues at 5Å or less from a ligand included in the crystallographic structure were considered part of the protein interface.

Results

Our results on the MolMov dataset showed statistically significant predictions when all the residues predicted by JAMMING in 500 conformers of every protein were considered. JAMMING performed better than 3 web servers (cons-PPISP[9], meta-PPISP[10] and PINUP[11]) tested. Despite these results, JAMMING predicted many interface residues outside of the observed interface. We learned that at least 50% of these false-positive predictions were indeed part of other known interfaces for the proteins under study.

In the case of proteins in CASP9, we considered two additional aspects: i) the generation of a reliable model and ii) picking the correct interface residues predicted by JAMMING to match those included in the crystallographic data. For the first aspect, we decided to model only proteins having a clear homologue in a protein family (*i.e.*, to be part of a PFAM family). For the second aspect, we classified the known protein interfaces according to the nature of the ligand and the connectivity of each residue involved in the binding of every ligand.

19 out of 60 "Human" protein targets satisfied the PFAM criteria. Unfortunatelly, none of our interface residues classifications rendered significant results, so we opted to visually identify residues on the surface that were part of a clique or connected subgraph. This decision was based on the notion that known protein interfaces present residues within a connected subgraph or clique, and that such residues

tend to be "exposed". The term "exposed" here refers to those residues with more than 6 $Å^2$ of surface area exposed and that visually were on the surface of the model. 6 of the 19 predictions submitted to the CASP9 experiment by our group included at least 2 binding sites.

Our results indicate that residue centrality (JAMMING) renders better performance than current web servers aimed at predicting protein interfaces. A limitation of this approach is the false positive rate. However, automatic or supervised methods may be used to assist the prediction of relevant binding sites.

Availability

JAMMING is available at http://bis.ifc.unam.mx/jamming.

1. http://pfam.sanger.ac.uk/.
2. http://www.drive5.com/muscle/
3. http://www.salilab.org/modeller/
4. http://igs-server.cnrs-mrs.fr/elnemo/
5. http://bis.ifc.unam.mx/jamming/
6. http://www.molmovdb.org/
7. http://scop.mrc-lmb.cam.ac.uk/scop/
8. http://www.cs.waikato.ac.nz/ml/weka/
9. http://pipe.scs.fsu.edu/ppisp.html
10. http://pipe.scs.fsu.edu/meta-ppisp.html
11. http://sparks.informatics.iupui.edu/PINUP/

Jiang_Assembly

Protein Structure Prediction by a combination of the threading and fragment-based assembly method

Yun Hu, Aiping Wu, Liqing Tian,Wentao Dai, Taijiao Jiang
National Laboratory of Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China;
taijiao@moon.ibp.ac.cn

The structure prediction server, Jiang_Assembly combines our threading program FR-t5 and our de novo prediction method. For the easy and medium targets, a series of structure models were first generated by FR-t5, and the final ones were selected using a model selection protocol. For the hard targets without detectable good templates, their structures are predicted by our de novo prediction method.

Methods

We have developed a de novo prediction method based on fragment assembly strategy. First, the 9 residue fragment library was constructed by following Rosetta's procedure[1]. Second, the initial structure was constructed by extending one residue at a time from N-terminus to C-terminus according to structural compatibility between local fragments. Third, the conformation space was explored by Monte Carlo-guided fragment replacement according to a statistics scoring function to be described below, and a favorable conformation was obtained when the search was converged. The above processes were repeated at least one hundred times. Finally, the predicted topology of the main-chain was generated by clustering all the converged conformations.

The statistics scoring function used in the searching process is a five-bead coarse-grained (the main chain atoms of N, CA, C, O and the side-chain center) scoring function developed by our group (paper in preparation). This scoring function consists of an atom-atom contact potential, a hydrogen-binding term and a triple local conformational energy. Moreover, we have optimalized several factors to improve the performance of the fragment assembly strategy[1]. The most remarkable ones include: 1) consideration of the structural compatibility of adjacent structural elements; 2) use of 9 residue fragments for alpha-helical proteins; and 3) more candidate fragment templates for loop regions but less fragment templates for helix regions.

In CASP9, for each hard target the de novo strategy described above was used to create 100 structure models. The top5 models were selected based on the structure densities of SPICKER[2] clusters. For each easy and medium target, the model selection protocol operated as follows: let the maximal Z-score of the templates be [Z-score]max, up to 50 threading models are collected if their threading Z-scores were greater than [Z-score]max-1.0, these model were then evaluated by SELECTpro[3] scores, and the top ranked models were submitted.

Results

To assess the performance of the de novo method we developed, the method was applied to predict all 19 New Fold (NF) targets in CASP7. To ensure a fair comparison, the PDB database used to genrate fragment libraries and nr database used to generate sequence profiles were constructed based on the PDB database and nr sequence data available before the start of CASP7. As shown in table 1, it was estimated to rank ~5th by GDT_TS and ~7th by TM-score among all participated groups in CASP7.

The evaluation of the threading method was discussed with details in Jiang_THREADER Group.

Rank	GDT_TS		TMscore	
	Group	Score	Group	Score
1	Zhang	6.35	Zhang	6.48
2	Baker	6.31	Baker	6.41
3	SBC	6.10	SBC	6.15
4	CIRCLE-FAMS	5.88	CIRCLE-FAMS	5.95
5	**Our work**	**5.79**	Bates	5.84
6	GeneSilico	5.70	GeneSilico	5.83
7	MQAP-Consensus	5.67	**Our work**	**5.83**
8	Zhang-Server	5.66	SAM-T05	5.82
9	Bates	5.66	verify	5.77
10	SAM-T06	5.65	TASSER	5.75

Table 1. Performance of our method and other groups on 19 CASP7 New Fold targets

Availability

The web server based on our new cluster is under construction and will be available to public soon.

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**, 209-25.
2. Zhang,Y. & Skolnick,J. (2004). SPICKER: A clustering approach to identify near-native protein folds. *Journal of Computational Chemistry* **25**, 865-871.
3. Randall,A. & Baldi,P. (2008). SELECTpro: effective protein model selection using a structure-based energy function resistant to BLUNDERs. *Bmc Structural Biology* **8**:52

Jiang_THREADER

Protein Structure Prediction by FR-t5 threading method

Yun Hu, Lizong Deng, Taijiao Jiang
National Laboratory of Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China;
taijiao@moon.ibp.ac.cn

Jiang_THREADER is based on our recently developed threading program FR-t5 (Fold Recognition with 5 terms in the scoring function, paper in preparation). The philosophy of Jiang_THREADER is to assign a template for any query sequence.

Methods

We introduced a new type of information based on local structural preference of 3-residue and 9-residue fragments. By incorporating the new type of information with the widely used sequence profile, secondary structure information and hydrophobic scoring, we have developed a new threading method, called FR-t5. The alignment between the query and template was based on the dynamic programming. The protein structure models were built using MODELLER[1] based on the threading results.

Three template libraries were generated for CASP9 structure prediction. The first two libraries include PDB structures with sequence identity <= 70% and <= 90% respectively. The third library contains all 146726 single-chain structures from PDB. For each target, its sequence was aligned to each sequence of the template library and all the alignments were ranked based on their Z-scores. Based on Z-scores, the targets can be classified as easy (Z-score > =6.5), medium (4.0 <=Z-score <6.5) and hard (Z-score < 4.0) targets. For the easy and medium targets, top 5 Z-score models with their templates in the third, second libraries are submitted, respectively. For hard targets, their structure models were generated as follows: first, top 50 Z-score models were selected based on the first template library, then two model assessment scores were computed: (1) the agreement between the predicted second structures of PSIPRED[2], and the real second structure, and (2) the agreement between the contact patterns predicted by the SVMSEQ[3] program and the model. The models were then ranked by a linear combination of the Z-scores of the two assessment scores.

Results

In testing of the alignment accuracy on the SALIGN dataset, FR-t5 achieved an alignment accuracy of 58.9%. In testing the fold recognition sensitivity based on the Lindahl benchmark, FR-t5 recognized 84.0% (90.2%), 54.0% (71.9%), 35.0% (65.5%) of the Top1 (Top5) hits at the family, superfamily, and fold level, respectively. FR_t5 was compared with other methods (see Table 1 and 2).

Table 1. The alignment accuracy (%) on SALIGN

Methods	Acc
FR_t5	58.9
BLAST[a]	26.1
COMPASS[a]	43.2
SALIGN[a]	56.4
SPARKS[a]	53.1
SP3[a]	56.3
UNI-FOLD[a]	57.4

[a] Results are cited from Ref[4].

Table 2. Comparing FR-t5 method with other methods for fold recognition on the Lindahl benchmark

[a, b] Results are cited from from Ref[5]and Ref[6] ,respectively.

Methods	Family (%)		Superfamily (%)		Fold (%)	
	Top1	Top5	Top1	Top5	Top1	Top5
FR_t5	84.0	90.2*	54.0	71.9*	35.0	65.5*
FUGUE[a]	82.2	85.8	41.9	53.2	12.5	26.8
RAPTOR[a]	75.2	77.8	39.3	50.0	25.4	45.1
SPARKS[a]	81.6	88.1	52.5	69.1	24.3	47.7
FOLDpro[a]	85.0*	89.9	55.5	70.0	26.5	48.3
HHpred[b]	82.9	87.1	58.8	70.0	25.2	39.4
SP3[b]	81.6	86.8	55.3	67.7	28.7	47.4
SP4[b]	80.9	86.3	57.8	68.9	30.8	53.6
SP5[b]	81.6	87.0	59.9*	70.2	37.4*	58.6

*The best results are denoted by asterisk.

Availability

The web server based on our new cluster is under construction and will be available to public soon.

1. Sali,A. & Blundell,T.L. (1993). Comparative Protein Modeling by Satisfaction of Spatial Restraints. *Journal of Molecular Biology* **234**, 779-815.
2. Jones,D.T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* **292**, 195-202.
3. Wu,S. & Zhang,Y. (2008). A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. *Bioinformatics* **24**, 924-931.
4. Poleksic,A. & Fienup,M. (2008). Optimizing the size of the sequence profiles to increase the accuracy of protein sequence alignments generated by profile-profile algorithms. *Bioinformatics* **24**, 1145-1153.
5. Cheng,J. & Baldi,P. (2006). A machine learning information retrieval approach to protein fold recognition. *Bioinformatics* **22**, 1456-63.
6. Zhang,W., Liu,S. & Zhou,Y.Q. (2008). SP5: Improving Protein Fold Recognition by Using Torsion Angle Profiles and Profile-Based Gap Penalty Model. *PLoS One* **3**, e2325.

Jones-UCL

Protein fold and function prediction using pGenTHREADER and FRAGFOLD

D.W. Buchan[1], D. Cozzetto[1], S.M. Ward[1] and D.T. Jones[1]

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

d.jones@cs.ucl.ac.uk

URL: http://bioinf.cs.ucl.ac.uk

The Jones-UCL group's main efforts in CASP9 were in improvements to our fragment assembly method (FRAGFOLD [1]) and attempts at binding site prediction using a range of both in house tools and external methods. Most of our efforts were aimed at harder targets with a simple meta prediction method being used for target domains with obvious matches to template structures.

Methods

For CASP9 target domains which we believed could not be reliably predicted using fold recognition methods, FRAGFOLD was used to generate up to 5 structures. This approach to protein tertiary structure prediction is based on the assembly of recognized supersecondary structural fragments taken from highly resolved protein structures using a simulated annealing algorithm. FRAGFOLD v4.6 differs from previous versions mainly in the areas of improved long-range hydrogen-bonding, a new stochastic search procedure and improved fragment selection. More importantly we have re-tuned every adjustable parameter by running benchmarks on 70 small proteins of known structure. A few experimental options were tested for the first time e.g. building of multichain models. As many as 10000 structures were generated for each target domain using UCL's Legion supercomputer, 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 binding site prediction. our method is a semi-automated strategy for the prediction of binding site residues, which utilises the consensus of contact residues between homologous protein structure templates and their biologically relevant ligands. Initially, we calculate high confidence template structures and alignments to the CASP target sequences using pGenTHREADER [2]. If we only found remote, low confidence templates, we obtained the alignment to the target sequence from the DaliLite superposition[3] of its structure and our manually generated 3D model.

We then identified each template structure's ligand interacting residues. Using the annotations in SwissProt/Uniprot [4], in the Binding MOAD database [5] and in the literature. we identified the set of biologically valid ligands bound to each template. The residues mediating these interactions were then extracted from PDBSum [6]. To expand the list of putative binding residues, active site locations were also collected from the Catalytic Site Atlas [7].

Using the initial pair-wise alignments, the interacting residue coordinates were mapped from each template onto the target sequence. Consensus contact positions in the target were then calculated via a majority rule approach. Finally, the list of potential ligand binding residues was manually checked and modified according to complementary information derived from sequence analysis and predicted target structures.

Results

Predictions of folds were submitted for all targets and binding site predictions for all targets with a template match and for which a suitable ligand was known.

References

1. Jones D.T. (1997) Successful ab initio prediction of the tertiary structure of NK-Lysin using multiple sequences and recognized supersecondary structural motifs. PROTEINS. Suppl. 1, 185-191.
2. Lobley, A., Sadowski, M.I. and Jones, D.T. (2009) pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination, Bioinformatics, 25, 1761-1767.
3. Holm, L. and Rosenstrom, P. (2010) Dali server: conservation mapping in 3D, Nucleic Acids Res, 38 Suppl, W545-549.
4. The Uniprot Consortium (2010) The Universal Protein Resource (UniProt) in 2010, Nucleic Acids Res, 38, D142-148.
5. Benson, M.L., Smith, R.D., Khazanov, N.A., Dimcheff, B., Beaver, J., Dresslar, P., Nerothin, J. and Carlson, H.A. (2008) Binding MOAD, a high-quality protein-ligand database, Nucleic Acids Res, 36, D674-678.
6. Laskowski, R.A. (2009) PDBsum new things, Nucleic Acids Res, 37, D355-359.
7. Porter, C.T., Bartlett, G.J. and Thornton, J.M. (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data, Nucleic Acids Res, 32, D129-133.

Jscslb

A basic pipeline with manual input from stuctural alphabet prediction

O. Zimmermann[1], S. Mohanty[1] and J. Meinke[1]
1 – Simulation Laboratory Biology, Juelich Supercomputing Centre (JSC), Research Centre Juelich, 52425 Juelich, Germany
olav.zimmermann@fz-juelich.de

The purpose of our participation in CASP9 was testing some of our methods in the context of a full structure prediction pipeline. The outcome will serve as a guideline for the design of a fully automated method.

Methods

We first implemented a very basic workflow consisting of PsiBlast[1], mafft[2], hhpred[3] and Modeller[4]. For homology modeling we used PsiBlast hits realigned by the mafft-linsi algorithm. For fold recognition targets (i.e. no PsiBlast hit to any with an e-value $<10^{-3}$ we used hhpred. The respective target – template alignments were used as input for Modeller. No additional secondary structure prediction constraints for unaligned regions were used. Most parts of the pipeline have been automated using Biopython[5] for scripting.
Manual work was mainly involved in:

a) selection of fold recognition templates
b) target template alignment refinement
c) setup and model selection for template free modeling
d) model ranking

For a) and b) we used the correlation between the prediction from our structural alphabet prediction program LOCUSTRA[6] and the observed secondary structure of the template as well as a simple local profile alignment.
For each alignment variant 40-100 models were generated by Modeller.

For c) we used a development version of our Monte Carlo simulation software PROFASI[7] that features constraint guided simulation, so that dihedral constraints from LOCUSTRA could be used in template free modeling.

For d) we mainly relied on the normalized DOPE score from Modeller. Only for few targets manually adjustments were made to maximize the number of templates represented in the submitted models, or to include models with higher secondary structure content than the models with highest DOPE score. For the *ab initio* models the lowest energy structures with high secondary structure contents were selected.

Most targets were processed within one day. *Ab initio* simulations using PROFASI were allowed to run up to 72 hours.

Results

Preliminary results from those targets already released to the PDB and the models from the automated servers indicate that our very basic approach is on par with the average server performance (median rank 28). These results are based on a sequence independent alignment to the unsplit target structures using raw scores from Matt[8]. We noted in particular that:

a) template free modeling using *ab initio* Monte Carlo simulation with Profasi failed (T0531, T0564, T0569).

b) even for most targets with close homologues as templates the PsiBlast/MaFFT multiple alignments were inferior to the hhpred profile alignments.

c) correspondence of observed secondary structure to the LOCUSTRA predictions and local profile alignment improved the models for the few targets we tried it. (e.g. T0579, T0581, T0584, T0596)

d) The DOPE-score was a good measure for ranking models.

e) an undetected error in our submission script caused the wrong sequence to be submitted for all models that are incomplete at the N-terminus as those have been "threaded" to the first residue of the CASP-template.

Availability

After thorough analysis of the CASP9 results we plan to make an improved and fully automated version of our pipeline available.

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. Katoh,K., Kuma, K., Toh, H. & Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucl. Acids Res.* ***33,*** *511-518.*
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.* **234**, 779-815.
5. Cock PJ, Antao T, Chang JT, *et al.* . Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25,** 1422–1423.
6. Zimmermann, O. & Hansmann, U.H.E. (2008). LOCUSTRA: Prediction of local protein structure using a two-layer Support Vector Machine approach. *J. Chem. Inf. Model.* **48,** 1903-1908.
7. Menke,M., Berger,B. & Cowen,L. (2008). Matt: Local Flexibility Aids Protein Multiple Structure Alignment. *PLoS Comput Biol* **4,** e10.
8. Irback,A. & Mohanty,S. (2006). PROFASI: A Monte Carlo simulation package for protein folding and aggregation. *J. Comput. Chem.* **27**, 1548-1555.

Keasar

Refinement of server models by energy optimization

C. Keasar[1]
[1] – *Ben-Gurion University of the Negev*
Chen.keasar@gmail.com

Structural similarity between homologous proteins is the cornerstone of template based modeling (TBM), the most successful approach to protein structure prediction. However, homologous proteins are not structurally identical, and their dissimilarity sets an inherent limit to the accuracy of TBM. In practice, even this limit is typically not reached. Modelers often fail to identify the best possible templates and to optimally align them to the target. Bridging the inherent gap between the evolutionary inferred model and the native structure is the focus of model refinement, which uses chemical knowledge after all evolutionary evidence has been used. Thus, this final stage of TBM is essential. It may in principle lead to experimental quality models, the outmost goal of the protein structure prediction. However, model refinement turns out to be a hard task and the current state-of-the-art is that refinement typically fails to improve model quality. Considering the rapid pace at which structural genomics centers pour new potential templates into the protein data bank, one may speculate that the relative importance of model refinement is likely to increase over the next years.

Novel energy functions that capture elaborate features of native structures may considerably improve our ability to refine template based models. Specifically, our group focuses on the development of cooperative energy terms that try to impose native-like patters. We used the CASP platform to benchmark the usability of several new energy terms in the context of model refinement.

Methods

Energy function: all simulations used an energy function that includes non-cooperative torsion angle[1] and atom-pair potentials[2], and cooperative meta-terms that bound the latter from reaching values that are too low[4]. In addition the energy function includes a cooperative hydrogen bonding term[3], and a cooperative solvation term[5].

Refinement protocols: For the human prediction section we applied the following protocol: (I) server models were downloaded from the CASP website and energy minimized; (II) the minimized models were ranked by a weighted sum of the various energy values; (III) Highest ranking models were visually inspected; (IV) A few selected models were optimized by Monte Carlo minimization; and (V) final models were selected and ordered by energy and visual inspection. Each submitted model includes a reference to the original server model that it tried to refine. For the refinement section we manually manipulated the starting models and then applied energy optimization.

Self assessment: the results presented below are based on the native structures that were available on Sept17th 2010, and on the domain definitions suggested by the Zhang group[6]

Results

What went right: around 40% of the models that we submitted as human prediction are better than the original server models in terms GDT_TS. Improvement rates are even higher (~60%) in terms of RMS, however the interpretation of this result is somewhat unclear in high RMS models. These results are consistent with recent refinement experiments that use other decoy sets as starting points.

What went wrong: First, the improvements are rather minor, ranging from 0.1% to 6%. Second, the human intervention in the modeling that we did for the refinement section turned out to be a bad decision. None of the models were improved by the manual manipulation.

Availability

All the software that was used in this work is freely available at http://www.cs.bgu.ac.il/~meshi.

1. Amir ED, Kalisman N, Keasar C. Differentiable, multi-dimensional, knowledge-based energy terms for torsion angle probabilities and propensities. *Proteins: Structure, Function, and Bioinformatics*. 2008;72(1):62-73.
2. Summa CM, Levitt M. Near-native structure refinement using in vacuo energy minimization. *Proc. Natl. Acad. Sci. U.S.A.* 2007;104(9):3177-3182.
3. Levy-Moonshine A, Amir ED, Keasar C. Enhancement of beta-sheet assembly by cooperative hydrogen bonds potential. *Bioinformatics*. 2009;25(20):2639-2645.
4. Maximova T, Keasar C. unpublished
5. Kalisman N, Keasar C. unpublished
6. http://zhanglab.ccmb.med.umich.edu/casp9/native.html

KKE

Residue-residue contact prediction using predicted structure information

Shunta Kimura, Kei Kobayashi and Teppei Ebina
Department of Lifescience and Biotechnology, Tokyo University of Agriculture and Technology (TUAT)

We developed a residue-residue contact prediction method by selecting statistically possible structures from the prediction results of recently available structure prediction servers, such as Rosetta, Zhang server and MULTICOM. Our prediction was observed from the residue-residue contact profile (RRCP) of the highest ranked predicted structure.

The predicted structures of a target protein were ranked using only RRCP information when the similar predicted structures were obtained from the servers (total RMSD of the residue-residue distance profile among the predictions less than 30.0), and were ranked using RRCP information and our newly developed structure scoring method when the predicted structures were significantly different each other. An RRCP was scored higher when it contained more conserved residue-residue contacts among the predictions than other RRCPs.

The structure scoring method was performed as follows: we constructed a three-residue fragment dataset from 7769 protein domains listed in the PDB. The protein sequences were selected by PDB-REPRDB (http://mbs.cbrc.jp/pdbreprdb-cgi/reprdb_menuJ.pl) with sequence identity, RMSD and structure resolution threshold of 40%, 4À and 2.5À, respectively. This yielded a total of 583,344 three-residue fragments, which represented an average of 73 fragments for each three residue fragment ($8000=20^3$ three-residue fragment species). We calculated the dihedral angle, the secondary structure and accessible surface area using DSSP. Then, a three-residue fragment in a target structure was scored as $\text{Score} = 1/(\text{Score}_{\Phi}\times\text{Score}_{\Psi}\times\text{Score}_{\text{ACC}),}$ where Score_{Φ}, Score_{Ψ} and $\text{Score}_{\text{ACC}}$ indicated the frequency rate of the target fragment with the values of Φ, Ψ and ACC, respectively. This score become higher when a target fragment formed unusual dihedral angle and existed in unusual environment of the respective three-residue fragment in the dataset. Finally, we selected a predicted structure with higher total fragment score and more conserved RRCP than those of others.

KnowMIN

Combined effect of Knowledge- and Physics-Based Potentials

Gaurav Chopra and Michael Levitt
Dept. of Structural Biology, School of Medicine, Stanford University
gaurav.chopra@stanford.edu

Knowledge-based potentials in various forms have been successfully used for protein structure predictions at previous CASP experiments, where they compared favorably to various physics-based potentials. With much recent advances in physics-based potentials[1; 2], it is timely to revisit the physics-based potentials and test the effect of combining them with the knowledge-based potentials for protein structure refinement. In CASP9 we submitted predictions for all the targets by processing them through three different pipelines. The template-based pipeline processed all the targets with PSI-BLAST[3] E-value below 0.01 or 3D-Jury[4] score above 45. The template-free modeling pipeline processed all the other targets. We tested our consistent knowledge-based refinement protocol[5] for structure prediction. In addition, a combination of knowledge-based and various physics-based potentials were tested on the refinement targets. All steps in these pipelines can be automated: we hope to run them under the server category at future CASPs.

Methods.

For template-based and template-free modeling category, we used energy minimization on the server models with the knowledge-based (KB_0.1) potential[6] in ENCAD, the MESHI[7] force field and a combination of both KB_0.1 and MESHI. This pipeline has been tested extensively on all human and server models predicted in CASP7 and performed consistently well[5]. Further tuning of the pipeline was done based on the predicted secondary structure by psipred[8] and an approximation of GDT-TS score for the target sequence. We calculated approximate GDT-TS score using the 3D-Jury score and the length of the protein (GDT-TS score = 69*[3D-Jury Score / Number of Residues] + 16). Based on our consistency test on CASP7 models, we used KB_0.1 and MESHI refinement protocol[5] on the server models with all helical and loop residues for both template-based and template-free modeling category, defined by our approximate GDT-TS score. Server models with no alpha-helix were refined using MESHI minimization only for the target with approximate GDT-TS between 20% and 60%; all other beta proteins not in this range were minimized with the KB_0.1 and MESHI refinement protocol. Moreover, all other targets were also energy minimized with KB_0.1 and MESHI protocol for the comparative modeling category with GDT-TS score of 50% to 80%. Finally, the selection and ranking of these models was based on the KB_0.1 energy scores of initial and final energy minimized server models.

For the refinement targets, we used energy minimization with the KB_0.1 and MESHI refinement protocol as well as a combination of this protocol various state of the art physics-based potentials in GBSA implicit solvent[9]. We used AMBER99SB[10], CHARMM27-CMAP[11] and OPLSAA[12] potentials in GBSA implicit solvent combined with KB_0.1 and MESHI refinement protocol[5] for refinement. The physics based calculations used GROMACS[13-15] and then the models were process by ENCAD and MESHI softwares for KB_0.1 and MESHI refinement protocol. The OPLSAA using GBSA implicit solvent was tested extensively and performed well on a large dataset of decoys[16], but combinations of physics-based and knowledge-based protocol are largely untested. For refinement, we used the starting model provided and did not use the information about problematic regions in this model. Our protocol was applied on the entire structure with no sub-division into individual domains, even when the target was known to have two or more domains. These choices were made to provide a consistency check of our refinement protocol. Finally, the ranking of the refined models was based on their KB_0.1 energy scores.

Availability

The KB_0.1 and MESHI protocol[5] is available as an online server at http://csb.stanford.edu/koba_stable/. It calculates C□ RMS, GDT-TS and GDT-HA scores to a reference structure if given and to the starting model are calculated if no reference is given. We plan to release all successful prediction pipelines as online servers in the future and hope to run them under the server category at all future CASPs.

1. Wickstrom,L., Okur,A., Simmerling,C. (2009). Evaluating the Performance of the ff99SB Force Field Based on NMR Scalar Coupling Data. *Biophysical Journal.* **97**, 853-856.
2. Buck,M., Bouguet-Bonnet,S., Pastor,R.W., MacKerell,A.D.,Jr. (2006). Importance of the CMAP correction to the CHARMM22 protein force field: dynamics of hen lysozyme. *Biophysical Journal.* **90**, L36-38.
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, and Rychlewski L (2003) 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics*. **19**:1015-1018.
5. Chopra,G., Kalisman,N., Levitt,M. (2010). Consistent refinement of submitted models at CASP using a knowledge-based potential. *Proteins.* **78**, 2668-2678.
6. Summa, C. M., Levitt, M. (2007) Near-native Structure Refinement Using in Vacuo Energy Minimization. *Proc. Natl. Acad. Sci. U.S.A.* **104**:3177-3182.
7. Kalisman, N., Levi, A., Maximova, T., Reshef, D., Zafriri-Lynn, S., Gleyzer, Y., Keasar, C. (2005) MESHI: a New Library of Java Classes for Molecular Modeling. *Bioinformatics*. **21**:3931-3932.
8. 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.
9. Qiu,Q., Shenkin,P.S., Hollinger,F.P., 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. Hornak,V., Abel,R., Okur,A., Strockbine,B., Roitberg,A., Simmerling,C. (2006). Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins.* **65**, 712–725.
11. Mackerell,A.D.,Jr., Feig,M., Brooks,C.L.,III. (2004). Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J Comput Chem*. **25**, 1400–1415.
12. Kaminski,G.A., Friesner,R.A., Tirado-Rives,J., Jorgensen,W.L. (2001) Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on Peptides. *J Phys Chem B.* **105**, 6474–6487.
13. Hess,B., Kutzner,C., van der Spoel,D., Lindahl,E. (2008) GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comp.* **4**, 435-447.
14. Larsson,P., Lindahl,E. (2010) A High-Performance Parallel-Generalized Born Implementation Enabled by Tabulated Interaction Rescaling. *J. Comp. Chem.* **31**, 2593–2600.
15. Bjelkmar,P., Larsson,P., Cuendet,M.A., Hess,B., Lindahl,E. (2010). Implementation of the CHARMM force field in GROMACS: Analysis of protein stability effects from correction maps, virtual interaction sites, and water models. *J. Chem. Theory Comp.* **6**, 459–466.
16. Chopra,G., Summa,C.M., Levitt,M. (2008). Solvent dramatically affects protein structure refinement. *Proc Natl Acad Sci USA.* **105**, 20239–20244.

Kochanczyk

Active Site Prediction From Central Distances of Amino Acids

M. Kochanczyk[1]
[1] *Jagiellonian University, Krakow*
marek.kochanczyk@uj.edu.pl

Recently, in search for a general measure of solvent exposure, it was showed that distributions of central distances of amino acids (i.e. their distances to the center of the macromolecule) provide an elegant framework for the description of the hydrophobicity in (globular) proteins[1]. As it turns out that global analysis of hydrophobicity may be employed for the task of prediction of residues creating binding/active sites[2], in the confluence of these two ideas a hydrophobicity-probability-based method was developed and applied for several structures available for human predictors from server-generated previews in CASP9.

Methods

In our approach, we simply search for residues that occur in unusual distances to the geometric center of the protein. A mixture model of (radial and spherically symmetric) probability density functions, expressing the distribution of atoms as a function of the distance to the center of the molecule normalized bythe radius of gyration, is applied to quantify the degree of unexpectedness. Contributions of the mixture depend on the amino acid composition of a considered protein. The method is not yet fully automated, but at the current stage it seems to be helpful in visual inspection of globular proteins. With its high sensitivity but low specifity it could potentially be complemented with another technique that explores local characteristics of the structure.

Availability

Web server SurpResi for the prediction of functionally important sites based on unusual central distances of atoms is available at www.bioinformatics.org/surpresi. The input of the server is a file in the PDB format. The output is a downloadable PDB file, where beta factors are replaced by values that are inversely proportional to the probability of encountering a residue according to the model; a hierarchy of putative active site residues is included in the remark section.

1. Gomes, A.L., de Rezende, J.R., Pereira de Araujo, A.F., Shakhnovich, E.I. (2007). Description of atomic burials in compact globular proteins by Fermi-Dirac probability distributions. *Proteins* **66**(2), 304-20.
2. Brylinski, M., Prymula, K., Jurkowski, W., Kochanczyk, M., Stawowczyk, E., Konieczny, L., Roterman, I. (2007). Prediction of functional sites based o nthe fuzzy oil drop model. *PLoS Comput. Biol*.**3**(5):e94.

Kurcinski-Kihara

Protein structure prediction aided by global and local model quality assessment

M. Kurcinski and D. Kihara
[1] Department of Biological Sciences, [2] Computer Science, Purdue University
dkihara@purdue.edu

We submitted predictions in two categories: tertiary structure prediction (TS) and quality assessment of models in general (QA MODE 1). We submitted total 305 of TS models, 5 models per each of 61 targets from human-server set. In QA category 58 predictions were submitted.

Methods

SubAqua [1] - real-value model quality assessment method was combined with CABS [2], a reduced protein model aimed at prediction/refinement of protein models and investigating dynamics.

SubAqua predicts the quality of protein models, by combining two distinct scores, SPAD and Verify3D. The SPAD score [3] reflects the alignment stability and is obtained by generating suboptimal target-template alignments. Verify3D analyzes the compatibility of the model with its own amino acid sequence by considering structural environments of residues. SubAqua provides prediction of both global (RMSD and lga) and local (Ca atom displacement from its native location) model quality measures.

CABS predicts protein structures by highly efficient sampling of the molecule's conformational space [4]. At the same time it attempts to satisfy spatial restraints derived from the structure of the template. For the purpose of the current work CABS's algorithm has been modified to utilize both global and local quality measures of the starting models. Starting models are computed by threading methods and often referred to server models. Predicted global RMSD is used to adjust intensity of conformational sampling of the whole protein molecule, while local model quality measure is used to distribute unevenly sampling effort along the protein chain. Sampling is restricted in protein parts predicted to be correct and enhanced in those predicted to be wrong.

Results

SubAqua global quality predictions of 60 targets from human-server category were submitted. For 40 targets for which crystallographic structures has been already released, the Pearson's correlation coefficients has been calculated between SubAqua predictions and GDT-TM-score of assessed models, with average value equal to 0.39.

Evaluation of predicted tertiary structures of 40 out of 60 targets included calculation of RMSD, GDT-TS-score, and TM-score. There are 16 targets which have a model within RMSD of 6Å, 5targets within GDT-TS of 0.5, and 10 targets within TM-Score of 0.5, among the five submitted models.

Availability

SubAqua - http://kiharalab.org/SubAqua
CABS - http://www.biocomp.chem.uw.edu.pl/services.php

Acknowledgements

This work is supported by NSF (EF0850009, IIS0915801) and NIH (GM075004). Other support from NSF is also acknowledged (DMS80568).

1. Yang YD, Spratt P, Chen H, Park C & Kihara D. (2010). Sub-AQUA: real-value quality assessment of protein structure models. Protein Eng Des Sel. **23**(8), 617-32
2. Bowie JU, Lüthy R & Eisenberg D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. Science. **253**(5016), 164-70
3. Chen H, & Kihara D. (2008). Estimating quality of template-based protein structure models by alignment stability. Proteins 71, 1255-1274.
4. Kolinski, A. (2004). Protein modeling and strucutre prediction with a reduced representation. Acta Biochim Pol. **51**(2), 349-71.

LEE

Protein structure modeling by global optimization and dynamic fragment assembly.

Keehyoung Joo[1], Mina Oh[1], Juyong Lee[1,2], Jinhyuk Lee[1], Junsu Ko[1,2], Sangjin Sim[1], Hae-Seok Eo[1], Jinwoo Lee[3], Masaki Sasai[1,4], Chaok Seok[1,2], In-Ho Lee[1,5], and Jooyoung Lee[1*]

[1]*Center for In Silico Science, School of Computational Sciences, Korea Institute for Advanced Study, Seoul 130-722, Korea*

[2]*Department of Chemistry, Seoul National University, 151-747, Korea*

[3]*Department of Mathematics, Kwangwoon University, Seoul 139-701, Korea*

[4] *Department of Applied Physics, Nagoya University, Nagoya 464-8603, Japan*

[5]*Korea Research Institute of Standards and Science (KRISS), 305-600, Korea*

* jlee@kias.re.kr

Template Based Modeling (TBM): For TBM targets, we applied a similar procedure used in CASP7/8, where protein modeling was executed by three layers of global optimization[1]. In CASP9, a new procedure for automatic template combination was introduced. For a query sequence, FOLDFINDER[1] is used to identify top templates from PDB. Templates with almost identical structures (TM-score $\geq$ 0.975) are excluded. Likewise, dissimilar structures (TM-score <0.65) from the top template are also excluded. By combining a sequence-based score and a 3D-based score, we identified core templates and secondary optional templates. The number of core templates ranges from 1 and 12, and that of optional templates is between 0 and 4. Template combination generates a set of template lists, which includes all core templates and all possible combinations (up to 2^4=16) of optional templates. For each template list, we performed multiple sequence alignment[2], chain building[3], quality assessment (QA) and side-chain remodeling by successively applying the global optimization method, conformational space annealing (CSA) [4]. QA was used to select final models to submit. For oligomeric structure prediction, we used templates in oligomeric states to determine relative positions between 3D models. The server prediction of *gws* and human prediction of *LEE* applied the identical protocol described above except that additional templates from 3D-jury and HHsearch were considered in *LEE*. The initial two-week predictions by g*ws* were plagued by unnoticed incomplete template database missing best templates and/or using a wrong program without template restraints.

Free Modeling (FM): When proper templates were not identified, an *ab initio* protein modeling procedure was used to generate protein 3D models. The method is based on the conformational search by CSA and dynamic fragment assembly (DFA) [5]. The idea of DFA is to construct a restraint energy function which contains the information of local interaction from a 9-residue fragment library. The library is generated as in a typical fragment assembly method. DFA is combined with DFIRE statistical potential and various physics-based terms for proper stereochemistry of proteins.

Loop Modeling: During TBM, when loops are identified (gap regions of multiple sequence alignments, regions with irregular stereochemistry, etc.), they are re-modeled by DFA introduced above.

Consensus Modeling by LEEcon: This is to generate consensus models using SERVER predictions. We performed structural clustering of SERVER models and identify the largest cluster. The clustering was set so that about 20 models are in the largest cluster. Using all models in this cluster as templates, we followed the identical procedure of global optimization used in TBM.

Model Refinement: For refinement targets, first, we deleted inaccurate regions described in the remarks provided by CASP. After including the given models for refinement into the core templates of TBM, we followed the TBM procedure described above. Side-chains are re-built from scratch following the TBM protocol.

Acknowledgements: This work was supported by Creative Research Initiatives (Center for *in silico* Protein Science, 2009-0063610) of MEST/KOSEF.

1. Joo,K., Lee,J., Lee,S., Seo,J.-H., Lee,S.J. & Lee,J. (2007) High-accuracy template based modeling by global optimization. *Proteins*, **69**(S8), 83-89.
2. Joo,K., Lee,J., Kim,I., Lee,S.J. & Lee,J. (2008) Multiple sequence alignment by conformational space annealing. *Biophys J.*, **95(**10), 4813-4819.
3. Joo,K., Lee,J., Seo,J.-H., Lee,K., Kim,B.-G., & Lee,J., (2009) All-atom chain-building by optimizing MODELLER energy function using conformational space annealing, *Proteins*, **75**(4), 1010-1023.
4. Lee,J., Lee,I.-H., & Lee,J., (2003) Unbiased global optimization of Lennard Jones clusters for N <= 201 by conformational space annealing method, *Phys.Rev.Lett.*, **91**, 080201.
5. Sasaki,T., Cetin,H., & Sasai,M., (2008) A coarse-grained Langevin molecular dynamics approach to de novo protein structure prediction. *Biochemical and Biophysical Research Communications*, **369**, 500-506.

LenServer

De novo Prediction of Protein Backbone by Parallel Ant Colonies

X. Huang[1], H. Wu[1], J. Wu[1], S. Chen[1], D. Miao[1] and Q. Lü[1,2]
[1] - *School of Computer Science and Technology, Soochow University, China*
[2] - *Jiangsu Provincial Key Lab for Information Processing Technologies, China*
qiang@suda.edu.cn

We have developed a new parallel approach for *de novo* prediction of protein backbone. It can combine the different sources of energy functions by sharing one pheromone matrix based on ant colony optimization approach.

Methods

(1) In the first phase, we employed a fold guided fragments generator, which is based on our SVM protein fold predictor [1]. Apart from the original fragments generated by Robetta online server [2], we use our SVM protein fold predictor to recognize the SCOP fold which the target protein belongs to, and we rebuild the fragment database using the proteins which share the same fold with the target. So our prediction method is performed on such two types of fragment libraries. We submit the prediction results based on the original fragments as server prediction, and fold guided fragments as human prediction.

(2) In the second phase, we use our parallel ant colonies for fragments assembly. Each colony can search the best backbone with an energy function based on general ACO framework [3]. The parallel colonies use different energy scores, and cooperate each other with sharing the pheromone matrix which accumulates the search knowledge of each colony. In this way, the final backbones are jointly determined by all the energy functions adopted by parallel colonies. The energy functions used are the same as Rosetta *ab initio* protocol [4], named score0,1,2,5,3. Four colonies adopt score0,1,2,5 respectively, and the rest colonies adopt score3. All the colonies are running in parallel exchanging their search knowledge stored in pheromone matrix.

(3) The best backbone found by each colony forms a decoy set. We apply a cross operation on it simply by exchanging two domains of randomly selected two decoys. Therefore, we double the size of the decoy set in a very short time. The final decoys (usually 1000 conformations) are clustered by quality threshold clustering algorithm. The biggest 5 clusters are selected and the center of the clusters are submitted as model1-5.

Results

As we only focus on *de novo* prediction tasks, we report the results for those identities less than 30% found by BLAST. We found twelve such target domains from all CASP9 targets: T0531, T0534, T0537, T0547_3 (denoting domain 3 of T0547), T0547_4, T0564, T0571_1, T0578, T0581, T0604_3, T0618, and T0621 [5]. Because of the limitation of our computing power, the size of decoy is only about ~800. We apply 16 parallel colonies for each prediction.

As for server predictions, the TM-scores of model 1for T0534, T0581, and T0604_3 are 0.231, 0.299, and 0.201 respectively, while the best TM-scores from the decoys are 0.263, 0.351, and 0.202 respectively. The sizes for each of the decoys are 352, 960, and 16 respectively.

As for human predictions, the best TM-scores from the decoys for T0531, T564 and T578 are 0.306, 0.453 and 0.309 respectively. The sizes for each of the decoys are 1280, 960, and 640 respectively.

Availability

All the implementations can be accessed by online service at http://ckcst20.suda.edu.cn:8080/test.

Acknowledgement

Supported by the National Natural Science Foundation of China (NSFC) under grant No. 60970055.

1. Guo H., Lü Q., Wu H., Wu J., Yang P., & Huang X. (2010) A SVM classifier for protein fold recognition, China Journal of Bioinformatics (*in press*).
2. http://robetta.org/fragmentsubmit.jsp
3. Dorigo M., & Stützle T. (2004) Ant Colony Optimization, MIT press.
4. Rohl C.A., Strrauss C.E.M., Misura K.M.S., & Baker D. (2004) Protein structure prediction using rosetta, Methods in enzymology 383, 66-93.
5. http://zhanglab.ccmb.med.umich.edu/casp9, 2010-9-16

LOOPP

A server for sensitive detection of structural templates and homology modeling

Brinda Vallat, Thomas Blom, Baoqiang Cao, Ravikant Dintyala, Shruthi Vishwanath and Ron Elber
Institute for Computational Engineering and Sciences, University of Texas at Austin, Austin, TX
Email: ron@ices.utexas.edu

Methods

LOOPP is a homology modeling server. It is based on a template detection algorithm learned by mathematical programming techniques that combines a large number of signals and significantly enhances typical detection capabilities (PSI-BLAST) by about 50 percent. It also uses a novel algorithm for alignment, and it finally builds atomically detailed models with Modeller (using the identified templates and our alignments of the target sequence into them). We use a combination of decision trees that constitute a "forest" to identify templates and assess the models. Each branch of the decision tree is a mathematical programming model and the confidence levels of the decision trees decrease as we move down the forest. The strength of the algorithm is in the very large training and test sets that we develop and use. The algorithm is fast and takes (at most) a few hours to build about 20 models per proteins.

Availability

The LOOPP server may be used by submitting a query sequence at the following website:

http://clsb.ices.utexas.edu/loopp/web/

Additionally, LOOPP is fully open-source software. It is MPI-based software written to be platform agnostic, but is primarily tested on Linux-based clusters. Source-code (perl, c++, FORTRAN) is available via anonymous svn at

https://svn.ices.utexas.edu/repos/clsb/trunk/loopp

1. Brinda Kizhakke Vallat, Jaroslaw Pillardy, Peter Majek, Jaroslaw Meller, Thomas Blom, BaoQiang Cao, and Ron Elber, "*Building and assessing atomic models of proteins from structural templates: Learning and benchmarks*", Proteins: Structure, Function, and Bioinformatics, 76:930-945 (2009).
2. Brinda Kizhakke Vallat, Jaroslaw Pillardy, and Ron Elber, "*A template-finding algorithm and a comprehensive benchmark for homology modeling of proteins*" , Proteins: Structure, Function, and Bioinformatics, 72:910-928 (2008).
3. Octavian Teodorescu, Tamara Galor, Jaroslaw Pillardy, and Ron Elber, "*Enriching the sequence substitution matrix by structural information*", Proteins: Structure, Function and Bioinformatics, 54:41-48(2004)
4. Jaroslaw Meller and Ron Elber, "*Linear Optimization and a double Statistical Filter for protein threading protocols*", Proteins, Structure, Function and Genetics, 45,241-261(2001)
5. Dror Tobi and Ron Elber, "*Distance dependent, pair potential for protein folding: Results from linear optimization*", Proteins, Structure Function and Genetics, 41, 40-16 (2000).

Additional literature references are available at the LOOPP website given above.

Lovell_group

Distinguishing Functional and Structural Constraints on Evolution to Predict Binding Sites

S.G. Williams,[1] R.M. Ames,[1] Julia Handl[1], and S.C. Lovell[1]
[1] – *Faculty of Life Sciences, University of Manchester, UK*
simon.lovell@manchester.ac.uk

The identification of sites where proteins interact with other molecules is a useful first step in identifying likely function. In CASP 9 we developed a new combined method based in part on the work of others and in part on our own previous research. This method is based on two observations: (i) homologous proteins tend to bind ligands at similar sites and with similar modes and (ii) ligand binding residues will have different substitution patterns compared with similar residues in similar structural environments that are not binding ligands. The key to the second part is understanding and quantifying the background distribution, i.e. those evolutionary constraints that arise from protein structure.

Methods

We developed a combined method to predict binding sites. Firstly, we used models of protein structures generated by the I-TASSER[1] server. BLAST[2] was used to search target sequences against the PDB[3] and the NCBI non-redundant data set. Homologous structures with bound ligands were inspected to inform the choice of binding residues. In addition, sequences form the non-redundant data set were aligned with MUSCLE[4], and these multiple sequence alignments used as input for CRESCENDO[5].

CRESCENDO uses environment-specific substitution tables (ESSTs[6]) to quantify the degree of sequence conservation due to protein structure. ESSTs describe the differences in substitution patterns between local environments within a structure. Aspects of structure that contribute to evolutionary constrain, and hence substitution differences, are features such as solvent accessibility, secondary structure and hydrogen bonding. CRESCENDO identifies this structural conservation and corrects for it. Any additional conservation is due to constraints that are not accounted for. A major source of this constraint is from protein function.

Scores from CRESCENDO were smoothed in three dimensions, and combined with information from homologous structures with bound ligands. From this information binding sites were predicted by inspection.

Results

We made predictions of binding sites for all 127 CASP 9 targets. However, in 68 of these we had little confidence, and think that it is likely that no small-molecule binding site exists. A total of 44 predictions were made on the basis of the CRESCENDO score alone. Many of these appeared to have convincing binding sites, where several high-scoring residues clustered together in three dimensions. In addition a further 15 predictions were based on both CRESCENDO score and the identification of homologous proteins with bound ligands. It is these predictions, based on multiple sources of evidence, in which we have highest confidence.

Availability

http://www.bioinf.manchester.ac.uk/crescendo/

1. Zhang, Y. (2009) I-TASSER: Fully automated protein structure prediction in CASP8. *Proteins* **S9** 100-113.
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. Berman, H. M, Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E. (2000) The Protein Data Bank *Nucleic Acids Res.* **28**, 235-242
4. Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792-1797.
5. Chelliah, V., Chen, L., Blundell, T.L. and Lovell, S.C. (2004) Distinguishing Structural and Functional Restraints on Evolution in Order to Identify Interaction Sites J Mol Biol, **342**, 1487-504.
6. Overington, J., Donnelly, D., Johnson, M.S., Sali, A. and Blundell, T.B. (1992) Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Protein Sci* **1**, 216-26.

Mason

Predicting disorder and ligand-binding residues using a unified learning framework.

Huzefa Rangwala
Department of Computer Science, George Mason University
rangwala@cs.gmu.edu

To predict the local structure (disorder) and functional (ligand-binding) properties of protein residues we used a support vector machine based method. Our approach used sophisticated window-based integrative profile kernels that would capture information features from and around the residue of interest and couple it with heterogeneous information derived from varies other sequence-derived properties.

Methods

We used a support vector machine based tool called svmPRAT[2] to predict the local structure and function properties of individual protein residues. svmPRAT is one of the first tools developed to allow life science researchers to quickly and efficiently train SVM-based models for annotating protein residues with any desired property. svmPRAT can utilize any type of sequence information associated with the residues as well as information extracted from neighboring residues (local window concept). The program also implements a flexible window encoding scheme that differentially weighs information extracted from the neighboring residues based on their distance to the central residue.

svmPRAT was used to discriminate between residues belonging to ordered versus disordered regions. The feature sets were PSI-BLAST PSSM[2,] BLOSUM62 sequence features, and predicted secondary structure using svmPRAT itself. The window parameters of the base window kernel were varied to select the best model for prediction on the CASP 9 targets. Finally, linear, radial basis function, and a novel second order exponential kernel were tested on a smaller benchmark and used for the final prediction models. It was seen that the second order kernel along with use of the profile, BLOSUM and secondary structure features showed the best classification precision and recall and as such was selected for submission of the final disorder prediction during CASP 9.

We used similar models for predicting the ligand-binding residues but used only profile information in contrast to the disorder prediction models. We also followed a transfer learning methodology where the task from one domain could be used for supplementing a task in a related domain. We basically used the results from the disorder prediction models as well as secondary structure prediction models as additional features for the ligand-binding prediction models. This methodology when tested across a set of 500 proteins we observed that secondary structure proved to be a better source domain for the ligand-binding prediction dataset.

Results

In our previous work we have evaluated svmPRAT[2] on several classification and regression problems including disorder prediction, residue-wise contact order estimation, DNA-binding site prediction, and local structure alphabet prediction. svmPRAT has also been used for the development of state-of-the-art transmembrane helix prediction method called TOPTMH[3]. This toolkit developed provides practitioners an efficient and easy-to-use tool for a wide variety of annotation problems and was deployed as a backend for CASP 9.

Availability

http://www.cs.gmu.edu/~mlbio/svmprat

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. Rangwala H, Kauffman C & Karypis G (2009). svmPRAT: SVM-based Protein Residue Annotation Toolkit. *BMC Bioinformatics* **10:** 439
3. Ahmed R, Rangwala H & Karypis G (2010) TOPTMH: Topology Predictor for Transmembrane Helices. *Journal of Bioinformatics and Computational Biology.* **8:1,** 39-57.

McGuffin

Manual Prediction of Tertiary Structure, Disorder and Binding Site Residues

D.B. Roche[1] S.J. Tetchner[1] and L.J. McGuffin[1]
[1] - *School of Biological Sciences, University of Reading, Whiteknights, Reading RG6 6AS, UK*
l.j.mcguffin@reading.ac.uk

Our manual predictions were for the most part automated and we used the same basic methods that were developed for the IntFOLD and ModFOLD servers[1] (see our server abstracts for more detail). However, we also made use of the 3D server models, we heavily relied on our QA results obtained from ModFOLDclust2[2] and we used a considerable amount of manual intervention for predicting binding site residues in an attempt to add value to our automated FunFOLD[3] predictions.

Methods

For the tertiary structure (TS) category, our manual predictions were made using ModFOLDclust2 for model selection. The top five 3D server models, ranked according the ModFOLDclust2 global quality scores, were selected and submitted as TS predictions. The only major modifications made to the models were in cases where the full backbone trace did not exist, in which case the program BBQ[4] was used to reconstruct the chain. In addition, for each model the ModFOLDclust2 predicted per-residue error was added into the B-factor column for each set of ATOM records.

Manual predictions were submitted in the disorder prediction (DR) category using the DISOclust method[5]. The same protocol was followed that was used for the IntFOLD-DR server predictions; however, in this case all 3D server models were used to supplement the limited selection available from the nFOLD4 method. For each target, the tarball containing all CASP9 server models was submitted to the IntFOLD server and the resulting DR output files were then uploaded using the CASP8 manual submission form.

We attempted to add value to our automated binding site residue (FN) predictions by using the standalone version of our new FunFOLD method along with better 3D server starting models and QA information from ModFOLDclust2. The results from the standalone version of FunFOLD were visually inspected and included in the prediction if they followed the subsequent criteria: 1. The global quality score for the start model was acceptable; 2. Model-to-template superpositions were good; 3. Residues were in contact with more than two well superposed ligands; 4. The predicted residues were not in a disorder region according to DISOclust. If no prediction was made by the initial FunFOLD run, template-to-model superpositions were also carried out in addition to model-to-template superpositions in order to increase the number of possible ligands. The manual intervention described above could be automated and we intend to include such restraints in the next version of FunFOLD.

Results

Early results indicate that our manual predictions for the FN and TS categories show an improvement over our respective server predictions. According to paired Wilcoxon signed rank sum test using MCC scores for 21 targets, the manual FN submissions are significantly better than our server FN predictions with a p-value of 0.0071. The manual DR predictions are more selective than those from our server although we can measure no significant difference in scores using the data that is presently available.

Availability

An alpha version of the IntFOLD server with graphical output is available at: http://www.reading.ac.uk/bioinf/IntFOLD/IntFOLD_form.html.

The ModFOLDclust2, DISOclust and FunFOLD software can be downloaded from: http://www.reading.ac.uk/bioinf/downloads/

1. McGuffin,L.J. (2008) The ModFOLD Server for the Quality Assessment of Protein Structural Models. *Bioinformatics.* **24**, 586-587.
2. McGuffin,L.J. & Roche,D.B. (2010) Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments. *Bioinformatics*. **26**, 182-188.
3. Roche,D.B., Tetchner,S.J. & McGuffin,L.J. (2010) FunFOLD: an improved automated method for the prediction of ligand binding residues using 3D models of proteins. *In preparation*.
4. Gront,D., Kmiecik,S., Kolinski,A. (2007) Backbone building from quadrilaterals: a fast and accurate algorithm for protein backbone reconstruction from alpha carbon coordinates. *J Comput. Chem*. **28**, 1593-1597.
5. McGuffin,L.J. (2008) Intrinsic disorder prediction from the analysis of multiple protein fold recognition models. *Bioinformatics*. **24**, 1798-1804.

MeDor

Philippe Lieutaud, Bruno Canard and Sonia Longhi
AFMB, UMR 6098, CNRS and Universities Aix-Marseille I and II, FRANCE

We have previously reported that combining various predictors can yield more accurate results (Bourhis et al. 2007; Ferron et al. 2006). .We therefore developed a web metaserver, referred to as MeDor, (http://www.vazymolo.org/MeDor/index.html) that allows fast, simultaneous disorder analysis of a query sequence by multiple disorder predictors and provides a graphical interface with a unified view of the outputs. It provides a global overview of various predictions relying on different philosophies, and considerably speeds up the disorder prediction (Lieutaud et al. 2008) (732 downloading on Sept 17, 2010).

As a next step to further speed up the analysis of protein disorder, we have implemented in MeDor four methods generating consensus predictions and used them to analyze CASP9 targets. MODEL 1 corresponds to the result of a consensus method computed with a weighting and a bonus system, in addition of which an automatic refinement of the boundaries, based on the HCA plot, was applied. MODEL 2 is the same as MODEL 1 except that no HCA-based boundary refinement was applied. MODEL 3 corresponds to the result of a consensus method computed with a different weighting philosophy. MODEL 4 corresponds to the result of a human analysis of the MeDor output. The bonus system takes into account the agreement among the various predictors, with the notion that a prediction is more reliable if different predictors relying on different physico-chemical principles converge. The weighting we applied was empiric and based on the significant experience that we gathered through the use of the different predictors integrated within MeDor. Evaluation of the performance of these four consensus within CASP9 will help in designating the best model. This latter will be chosen for subsequent implementation in future public releases of the MeDor program.

1. Ferron F, Longhi S*, Canard B and Karlin D. (2006) A practical overview of protein disorder prediction methods. Proteins 65, 1-14.
2. Bourhis JM, Canard B and Longhi S (2007) Predicting structural disorder and induced folding: from theoretical principles to practical applications. Curr Protein Pept Sci, 8, 135-49.
3. Lieutaud P., Canard B. and Longhi S (2008) MeDor: a metaserver for predicting protein disorder. BMC Genomics, Sep 16;9 Suppl 2:S25.

MeilerLab

Folding membrane proteins using sequence-independent templates

B.E. Weiner[1], N. Woetzel[1], M. Karakas[1] and J. Meiler[1]
[1] – *Vanderbilt University, Department of Chemistry, Nashville, TN, 465 21st Ave South BIOSCI MRBIII*
brian.weiner@vanderbilt.edu

Membrane proteins (MPs) remain challenging *de novo* structure prediction targets owing in large part to the scarcity of unique MP structures available in the PDB. The lack of suitable templates substantially hampers traditional template-based modeling approaches. In order to circumvent this problem, BCL::Fold, a template-free *de novo* protein structure prediction method, has been modified to fold MPs using sequence-independent templates. Templates are selected based on secondary structure composition rather than a sequence alignment. Templates can be recombined to create novel fold topologies.

Methods

BCL::Fold assembles discreet secondary structure elements (SSEs) using a Monte Carlo algorithm with a knowledge-based (KB) scoring function. The method was first adapted to incorporate MPs with the inclusion membrane-specific KB scores to assess the alignment of SSEs within the membrane, amino acid environments, and the radius of gyration.

A fold template library was generated from a non-redundant data set of proteins culled from the PISCES[1] server. The library contained both soluble and membrane proteins. Additional Monte Carlo moves were added to the algorithm to allow for SSE placements into complete or partial fold templates. Additionally, novel templates are generated by recombining partial templates throughout the assembly process.

The benchmark set of sixteen MPs was designed to sample varying protein sizes, topologies, and multimeric states. After secondary structure prediction, 50,000 models were generated using BCL::Fold. The top 10,000 models by score were clustered and cluster centers were compared with the native structure to evaluate prediction accuracy. In order to assess the impact of fold templates on the folding process, results were compared to structures generated without utilizing the template-based moves.

Results

Folding MPs using sequence-independent templates improved structure prediction accuracy, particularly for beta-barrel proteins. Overall, the method achieved moderate success in a difficult prediction area. The method is currently being expanded to incorporate experimental restraints to facilitate structure determination using sparse data.

1. Wang, G. & Dunbrack, R.L. (2003). PISCES: a protein sequence culling server. *Bioinformatics* **19**, 1589-1591.

MeilerLab

BCL::Fold – A novel de novo protein tertiary structure prediction method

M. Karakas[1], N. Woetzel[1], R.D. Starizbichler[1], N.S. Alexander[1], S. Lindert[1], J. Koehler[1] and J. Meiler[1]
[1] – *Vanderbilt University, Department of Chemistry, Nashville, TN, 465 21st Ave South BIOSCI MRBIII*
mert.karakas@vanderbilt.edu

Novel fold topologies not yet represented in the PDB are still found in large macromolecular complexes or membrane proteins. Structure determination is challenging for such proteins as many of these systems evade crystallization and alternative approaches such as cryo-electron microscopy or EPR spectroscopy yield low-resolution or sparse data sets. These proteins are typically too large for computational *de novo* structure prediction. The present algorithm seeks to expand the limits for de novo protein structure determination. It assembles predicted secondary structure elements (SSEs) in space, i.e. α-helices and β-strands, before adding the connecting loop regions or amino acid side chains. Size and complexity limits of previous approaches are overcome by discontinuing the amino acid chain in the folding simulation and limiting the sampling of flexible loop regions. Employing a Monte Carlo procedure, the sampling trajectory is guided by knowledge based potentials that evaluate amino acids' pair interaction and environment, SSE packing, loop closure, and protein compactness. The method is tailored to be used in conjunction with low-resolution or sparse experimental data sets.

Methods

Three secondary structure prediction methods, PSIPRED[2], PROFPHD[3] and JUFO[4], have been equally weighted to achieve a consensus three state secondary structure prediction. Stretches of sequence with consecutive α-helix or β-strand predictions above a given threshold were identified as α-helical and β-strand SSEs.

The predicted SSEs were then passed to the assembly protocol. The assembly method is a simulated annealing Monte Carlo minimization employing the Metropolis criterion. A variety of moves are utilized to generate new protein models throughout the minimization process, including SSE-based moves such as adding, removing, swapping, rotating, translating, flipping, bending as well as more advanced moves such as shuffling or flipping domains, twisting β-sheets, fixing β -sheet registers.

The assembly protocol was used to create 50,000 models. The best 10,000 models by energy were clustered. Clusters were calculated according to Cα RMSD100[6] using average linkage. The cluster centers of the twenty largest were added to a database of candidate structures. *Note that even if fold recognition methods identify a suitable template it is not employed by this method.* This experimental design was chosen to maximally leverage CASP for testing of the de novo folding algorithm. To assess if the algorithm folded any model with a native-like topology, even if it is not member of the twenty largest clusters, two models were added to the candidate structures: the model which was closest to the top ranked BIOINFO[5] homology model by Cα RMSD and the model with the best Z-score by mammoth[7] when compared to the PDB.

For each candidate model loops and side chains were added using Rosetta. The final five models for submission were selected as following: best by score after folding, best by score after refinement, lowest RMSD100 to homology model, best by mammoth Z-score, and one manually selected model.

1. Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology* **293**, 195-202.
2. B Rost, G Yachdav and J Liu (2004). The PredictProtein Server. Nucleic Acids Research 32(Web Server issue):W321-W326.
3. Meiler, J. & Baker, D. (2003). Coupled prediction of protein secondary and tertiary structure. *Proc Natl Acad Sci U S A* **100**, 12105-10.
4. Metropolis, N. R., A.; Rosenbluth, M.; Teller A. . (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087 - 1091.
5. Ginalski, K., Elofsson, A., Fischer, D. & Rychlewski, L. (2003). 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics* **19**, 1015-8.
6. Carugo, O. & Pongor, S. (2001). A normalized root-mean-square distance for comparing protein three-dimensional structures. *Protein Science* **10**, 1470-1473.
7. Ortiz, A. R., Strauss, C. E. M. & Olmea, O. (2002). MAMMOTH (Matching molecular models obtained from theory): An automated method for model comparison. *Protein Science* **11**, 2606-2621.

MidwayFolding

Automated Prediction Pipeline combining Homology-Based 'RaptorX' and Homology-Free 'ItFix' Tools for Local and Global Protein Structure

A.N.Adhikari[2], J. Hinshaw[2], J. Debartolo[1], J. Peng[6], J. Xu[6], M. Wilde[4], K. F. Freed[2,3,4] and T.R.Sosnick[1,4,5]

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

trsosnic@uchicago.edu

We have recently created an automated structure prediction pipeline for generating high accuracy structures by combining the strengths of template-based and free modeling methods. The pipeline begins with the template-based RaptorX[1] algorithm that first determines whether one or more template exists. If no templates exist, we use our free modeling ItFix/SPEED 2^o and 3^o structure prediction algorithm.[2] When the alignment contains insertions and deletions, local regions (e.g., large loops and ends) are described using our free modeling methods. This pipeline has been installed in several supercomputing architectures for high throughput structure prediction and was used for our group's submissions in the CASP9 experiment.

Methods

Global Structure Prediction: Our modeling tools include a Cbeta-level structure prediction algorithm, termed ItFix, that couples 2^o and 3^o prediction by iteratively fixing 2^o assignments of certain portions of the sequence after incorporating the influence of 3^o context. Our move set involves only changes in a single residue's φ/ψ backbone dihedral angles (i.e., not fragment insertion), with angles obtained from a PDB-based distribution appropriate for each amino acid, conditional on the type and conformation of the flanking residues. Furthermore, we use MSAs to enrich the φ/ψ sampling distribution in a manner not requiring structural knowledge of any protein sequence.[2] This method, termed "SPEED", removes the large impediment to accurate tertiary prediction that arises from the intrinsically low propensity of some residues to adopt the native dihedral angles. When the PSIPRED 2^o structure prediction for the target sequence yielded a very high confidence prediction at a given position for CASP9 targets, we fixed the conformational sampling at that position to retain the predicted 2^o structure, while allowing all other positions to search all 2^o structure types. The plethora of good models generated by this method enables us to cluster (hierarchical, by RMSD) and generate global and position resolved measures of confidence for the accuracy of the predictions. A major element in the successful predictions is our statistical potential, DOPE-PW, which includes only main chain heavy atoms and side chain Cbeta atoms. In addition to amino acid type, the statistical potential depends on 2^o structure and side chain orientation, thus yielding native-like structures with hydrophobic cores and hydrophilic surfaces. This energy function is minimized using a Monte Carlo Simulated Annealing (MCSA) algorithm using single pivot moves chosen at random positions in the sequence.

Local Structure Prediction: Often template-based protein structure prediction models are created by splicing together multiple structural fragments, which inevitably produces local regions that cannot be modeled based on known structures or folds. We used our homology-free tools in CASP9 to build these local regions, which very frequently are large loops or ends. Our loop/end modeling method generates

random conformations using the same single pivot (φ,ψ) move set as for structure prediction, and the interactions of the loop/end residues with each other and with the rest of the protein are computed to guide the conformational search, rather than building the fragment one residue at a time and trying to close the loop at the end. While some existing methods separate loop building and closure into two subsequent stages, our approach integrates the two in a single faster MCSA scheme, thereby retaining the tertiary context of the whole protein during the simulation attempting to find the best loop conformation. This tertiary context can be very important in longer loops for identifying crucial loop-protein interactions, thus greatly reducing the search space. The parallel algorithm generates possible loop conformations that are clustered and selected by a combined score using cluster tightness, DOPE-PW energy and solvent accessibility.

Availability http://sourceforge.net/projects/protlib/develop.

1. Xu, J., Li, M., Kim, D., and Xu, Y. (2003) RAPTOR: optimal protein threading by linear programming. J. Bioinform. Comp. Biol. 1, 95-117.
2. DeBartolo, J., Hocky G., Wilde M., Xu J., Freed K.F.,and Sosnick T.R. (2010). Protein Structure Prediction Enhanced with Evolutionary Diversity: SPEED., Prot. Sci. 19(3):520-34.

mn-fold

Ligand-binding Residue Prediction with LIBRUS in CASP9

Chris Kauffman[1] and George Karypis[1]
[1] – *University of Minnesota, Twin Cities*
kauffman@cs.umn.edu

We applied our previously developed method, LIBRUS[1], in CASP9 to predict residues likely to bind ligands.

Methods

LIBRUS combines two techniques which have become standard for protein prediction tasks: homology transfer and machine learning. Sequence alignment of the target against a database of known ligand-binding proteins is used to generate a homology transfer score (HTS) for each residue in the target. This score along with the target's PSI-BLAST profile and predicted secondary structure are fed into a support vector machine (SVM) model trained to identify binding residues. In our original work, we restricted the definition of binding residues to those forming contacts with large ligands (eight or more heavy atoms). For CASP9, we deployed two additional models: one to identify small ligands (fewer than eight heavy atoms) and one to identify all sizes of ligands.

Results

According to our internal evaluation, the original predictor, trained to identify ligands with eight or more heavy ligands at contact distance five Angstroms, was the most successful. Under those criteria for ligands and using 93 targets that had PDB files deposited, the predictor produced an area under the ROC curve of 0.67 and area under the precision/recall curve of 0.08 which corresponds to only 4% precision at 50% recall. Restricting the evaluation to only the 29 available targets which had ligands by this definition, performance was ROC=0.66 and precision/recall=0.24 which corresponds to 15% precision at 50% recall. Both performances are below the benchmark results.

Availability

LIBRUS models are available on request. Datasets and other tools associated with the original study may be downloaded from the web at http://bioinfo.cs.umn.edu/supplements/binf2009.

1. Kauffman,C., Karypis,G. (2009). LIBRUS: combined machine learning and homology information for sequence-based ligand-binding residue prediction. *Bioinformatics.* **25**, 3099-107

MOBI

MOBI: a web server to define and visualize structural mobility in NMR protein ensembles

Alberto J.M. Martin, Ian Walsh and Silvio C.E. Tosatto
1;*Alberto Jesus;Martin;Dr;Biocomp UP, Department of Biology, University of Padova, Italy*
2;Ian;Walsh;Dr;Biocomp UP, Department of Biology, University of Padova, Italy
3;Silvio C.E.;Tosatto;Proff;Biocomp UP, Department of Biology, University of Padova, Italy

MOBI [1] is a web server for the identification of structurally mobile regions in NMR protein ensembles. It provides a binary mobility definition that is analogous to the commonly used definition of intrinsic disorder in X-ray crystallographic structures. At least three different use cases can be envisaged: (i) Visualization of NMR mobility for structural analysis; (ii) definition of regions for reliable comparative modelling in protein structure prediction; (iii) definition of mobility in analogy to intrinsic disorder. MOBI uses structural superposition and local conformation differences to derive a robust binary mobility definition that is in excellent agreement with the manually curated definition used in the CASP8 experiment for intrinsic disorder in NMR structure. The output includes mobility-coloured PDB files, mobility plots and a FASTA formatted sequence file summarizing the mobility results.

The MOBI server and supplementary methods are available for non-commercial use at URL: http://protein.bio.unipd.it/mobi/.

1. Alberto J.M. Martin, Ian Walsh and Silvio C.E. Tosatto. MOBI: a web server to define and visualize structural mobility in NMR protein ensembles. Bioinformatics (2010) in press.

ModFOLDclustQ
ModFOLDclust2

Model Quality Assessment Using the ModFOLD Server

D.B. Roche[1] and L.J. McGuffin[1]
[1] - *School of Biological Sciences, University of Reading, Whiteknights, Reading RG6 6AS, UK*
l.j.mcguffin@reading.ac.uk

Predictions were submitted in the quality assessment (QA) category using the latest version of the ModFOLD server[1]. Here we describe two different clustering based methods, ModFOLDclustQ[2] and ModFOLDclust2[2], both of which were used to generate QA predictions in QMODE2 format. We have also developed a new method, ModFOLD 3.0, which is part of the new ModFOLD server and is able to operate in single-model mode. Further details concerning the ModFOLD 3.0 method can be found in our abstract describing the IntFOLD server.

Methods

ModFOLDclustQ

The ModFOLDclustQ method was developed in order to compare 3D models of proteins without the need for CPU intensive structural alignments[2]. The method is similar to our previous ModFOLDclust method[3,4], however a modified version of the structural alignment free Q-measure[5] is used instead of the TM-score[6] in order to carry out all-against-all pairwise model comparisons. The ModFOLDclustQ method has been rigorously benchmarked against the top established methods tested at CASP8[2].

ModFOLDclust2

The ModFOLDclust2 method was developed to provide increased prediction accuracy with minimal additional computational overhead. The global QA score from ModFOLDclust2 is simply the mean of the global QA scores obtained from the new ModFOLDclustQ method and the original ModFOLDclust method. The per-residue QA scores for ModFOLDclust2 were just taken directly from ModFOLDclust as no advantage was gained from combining the per-residue scores with those from ModFOLDclustQ.

Results

The ModFOLDclustQ method is competitive with the top clustering-based MQAPs that were tested at CASP8, for the prediction of global model quality. The ModFOLDclustQ method is also up to 150 times faster than the previous version of the ModFOLDclust method. In addition, a significant increase in accuracy can be obtained over the previous clustering-based MQAPs by combining the scores from ModFOLDclustQ and ModFOLDclust to generate the new ModFOLDclust2 method, with negligible impact on the total time taken to perform each prediction.

Availability

The latest version of the ModFOLD server (version 3.0 alpha) with graphical output is available at: http://www.reading.ac.uk/bioinf/ModFOLD/ModFOLD_form_3_0.html.

1. McGuffin,L.J. (2008) The ModFOLD Server for the Quality Assessment of Protein Structural Models. *Bioinformatics*.**24**, 586-587.
2. McGuffin,L.J. & Roche,D.B. (2010) Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments. *Bioinformatics*. **26**, 182-188.
3. McGuffin,L.J (2007) Benchmarking consensus model quality assessment for protein fold recognition. *BMC Bioinformatics*. **8**, 345.
4. McGuffin,L.J. (2009) Prediction of global and local model quality in CASP8 using the ModFOLD server. *Proteins*. **77**, 185-190.
5. Ben-David,M., Noivirt-Brik,O., Paz,A., Prilusky,J., Sussman,J.L. and Levy,Y. (2009) Assessment of CASP8 structure predictions for template free targets, *Proteins*, 77, 50-65
6. Zhang,Y. and Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins*. **57**, 702-710.

MQAPmulti
MQAPsingle
MetaMQAP
MetaMQAPclust

Model quality assessment using MQAPmulti

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

Four servers from our group participated in QA prediction in CASP9. Three servers operate on multiple models (MQAPmulti, MQAPsingle, MetaMQAPclust), the last one, MetaMQAP, is a trueMQAP, which is able to provide a useful score by considering only a single model.

Methods

MQAPmulti is a method recently developed in our laboratory, which evaluates the quality of many models generated for the target sequence. MQAPmulti compares structural features generated from a 3D model with those predicted from its primary sequence (secondary structure, solvent accessibility, contact maps), uses a DFIRE-like[1] statistical potential to estimate the value of pseudo-energy for a single model, uses hydrogen bonds pseudoenergy, and takes into account information from proteins that are evolutionary related to the target protein. In addition, MQAPmulti compares models using both GDT_TS[2], which is a method for scoring structural alignments that requires a rigid superposition of the models, and QS-score – our modification of Q-score[3] that works by estimating the structural relatedness between two protein structures based on comparison of intramolecular distances. Finally, MQAPmulti applies a method called "Correlation-Based Method for the Enhancement of Scoring Functions on Funnel-Shaped Energy Landscapes"[4] to combine a trueMQAP-like potential with clustering. MQAPmulti was optimized for the global model quality prediction in the following tasks: ranking models and selecting the best model from an ensemble of models.

Results

MQAPmulti was benchmarked using a CASP8 model dataset, which contains server models submitted for either single-domain targets or targets having all domains belonging to the same modelling category. The value of Pearson's correlation coefficient between MQAPmulti global score and the GDT_TS of models is 0.913, while the average value of the target-by-target correlation coefficients is equal to 0.953. Further, the average GDT_TS of the top ranked model by MQAPmulti is 0.692. To compare with, we present the same parameters computed for the CASP8 best performing methods in QA category: QMEANclust (0.885, 0.919, 0,674), ModFOLDclust (0.904, 0.925, 0.683), and PCONS (0.885, 0.924, 0,681).

Availability

MQAPmulti was under development during the first half of the CASP9 prediction season. The latest release of MQAPmulti was finished in July 2010. The MQAPmulti can be executed either as a standalone program or (from December 2010) as a web server.

Other methods

In contrast to MQAPmulti, which compares a model being scored with all models from an ensemble of models, MQAPsingle analyze a model against a subset of previously generated models, among which are ones created by the GeneSilico metaserver, Pcons, and HHpred methods.

MetaMQAP uses a machine learning approach to predict the deviation of C-alpha atoms of all residues in the model from their counterparts in the unknown native structure. This method combines the output from a number of primary MQAPs, including VERIFY3D, PROSA, BALA-SNAPP, ANOLEA, PROVE, TUNE, REFINER, PROQRES as well as local residue features: secondary structure agreement, solvent accessibility, residue depth. The global deviation is estimated in the form of RMSD and GDT_TS values.

MetaMQAPclust is a QMEANclust -like method. It first ranks all models according to the MetaMQAP score. Then, a 3D-Jury-like procedure is executed for 15% of the top-ranked models. The consensus score of a given model is its average GDT_TS to all models in the subset.

1. Zhou, H. & Zhou, Y. (2002). Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci. **11**, 2714–2726.
2. Zemla, A. (2003). LGA—a method for finding 3D similarities in protein structures. Nucleic Acids Res. **31**, 3370–3374.
3. Goldstein,R.A. et al. (1992) Optimal protein-folding codes from spin-glass theory. Proc.Natl Acad. Sci. USA. **89**, 4918–4922.
4. Stumpff-Kane, A.W., Feig, M. (2006). A correlation-based method for the enhancement of scoring functions on funnel-shaped energy landscapes. Protein Sci. **63**, 155–164.

MUFOLD
MUFOLD-server

Predicting Protein Tertiary Structure Based on a Multi-Dimensional Scaling Method

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

We have developed a system, MUFOLD[1], to predict tertiary structure from a protein sequence and cover both template-based and *de novo* structure predictions using the same framework.

Methods

The system includes three parts: 1) model construction using consensus constraints and a Multi-dimensional Scaling[2] (MDS) method, 2) model evaluation by a sampling-based machine-learning method, and 3) model refinement by combination of model and template information.

1) Model construction. The basic idea for model generation is to estimate the distances between backbone atoms for each pair of amino acid resides in the query sequence (i.e., distance constraints) and then calculate the backbone coordinates for all residues by applying MDS on the distance constraints. It has the following four key steps:

(a) to identify compatible templates and fragments of variable lengths in PDB (known structure) for a given query protein;
(b) to formulate pair-wise spatial constraints derived from the alignments between a query sequence and its templates and fragment hits;
(c) to check the consistency of the above spatial constraints to avoid conflicts;
(d) to apply MDS for generating initial 3D structural models from the spatial constraints.

When the query sequence and a template in PDB have significant sequence similarity, it is intuitive that the sequence alignment can be used to obtain high-quality distance constraints. Some of remote homologies can be obtained using sequence-profile alignment or profile-profile alignment, while many others cannot be obtained by sequence alignments. In these cases, we search for additional templates by threading alignments. We then build various distance constraints through sampling alignments and templates, and construct a model for each set of distance constraint through applying MDS. By MDS, MUFOLD can accommodate diverse spatial restraints retrieved from heterogeneous alignments including global or local alignments tailored for individual target.

While we can generate constraints for either the C-alpha atoms or all the backbone atoms, we use C-alpha atom distance constraints to generate initial model for computational efficiency, apply backbone atom distance constraints at the refinement stage, and generate full-atom side chains using Pulchra[3].

2) Model evaluation. Our sampling-based machine-learning method ranking structural models combines five existing knowledge-based scoring functions by a sampling method, and then uses radial basis function (RBF) neural networks to train a ranking function from the sampling distribution. In this way, the advantages of knowledge-based scoring functions, consensus approaches and machine learning-based scores can be combined. Our approach has two major contributions: (1) we integrate different features and scores systematically to obtain more discerning power for model quality assessment (QA); (2) we apply a sampling scheme and use sampling distribution features as input for more robust QA.

We have applied our method to two different datasets: one set is the CASP server prediction models of CASP8[4] targets and the other set includes the models generated by MUFOLD. The test results

show that our method performs significantly better than any of the five selected individual approach on both selection of top models and Spearman Correlation between the predicted GDT_TS scores and the actual GDT_TS scores of the models to their native structures.

3) Model refinement. An important feature of MUFOLD is to refine the models in two different ways. The first way is to refine the model by the combination of decoys and template information. In this case, we refine the restraints iteratively by combining the original restraints derived from the alignments (Dalign) and the measured distances from generated models (Dmodel) as: Drefine = λ*Dalign + (1 –λ)* Dmodel, $0 \leq \lambda \leq 1$. Here λ is set such that Drefine is the most consistent with Dalign and Dmodel, i.e., the sum of the distances from Drefine to Dalign and to Dmodel is the smallest. The second way is to refine the model based on consensus information. For this purpose, we evaluate and select top models by the ranking method described above, and then cluster these top models into groups. The models within the same group are similar and share some consensus distance constraints, which make it easy to derive a consensus model. By performing the refinement iteratively, the quality of models often improves while many deficiencies in the models are fixed over iterations. The current evaluation on the CASP9 targets with release structures shows that MUFOLD-Server performs much better than it did in CASP8.

We also use the strategy of MUFOLD-Server on CASP9 human prediction, where we only used all CASP9 server prediction models as templates. A preliminary assessment[5] of the CASP9 targets with release structures indicates that the method worked very well. For many targets such as T0569, T0576, T0582, T0592, T0606, T0618, T0622, the Model 1 of our human prediction is significantly better than the best model of all submitted CASP9 server predictions, and the average gain is 1.71, 3.33, and 2.64 points in terms of TM score, MaxSub score, and GDT-TS score, respectively

Availability

1. J. Zhang, Q. Wang, B. Barz, Z. He, I. Kosztin, Y. Shang, and D. Xu. MUFOLD: A New Solution for Protein 3D Structure Prediction. Proteins: Structure, Function, and Bioinformatics, 78, pp.1137-1152, 2009
2. I. Borg, P. Groenen, Modern Multidimensional Scaling – theory and applications, Springer-Verlag, New York, 1997.
3. M. Feig, P. Rotkiewicz, A. Kolinski, J. Skolnick, CL Brooks 3rd, Accurate reconstruction of all-atom protein representations from side-chain-based low-resolution models. Proteins: Struct Funct Bioinformatics 41(1), pp.86–97, 2000.
4. http://www.predictioncenter.org/casp8/index.cgi
5. http://sysbio.rnet.missouri.edu/casp9_assess/target_specific_mufold/

MUFOLD-MD

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

J. Zhang1, J. Zhang2,3, Q.Wang2, D. Xu2,3, Y. Shang2, I. Kosztin1
[1]*Department of Physics and Astronomy,* [2]*Department of Computer Science*
[3]*Christopher S. Bond Life Sciences Center, University of Missouri,*
Columbia, MO 65211, USA
KosztinI@missouri.edu

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

For CASP9, we have further developed our MUFOLD-MD server by implementing different prediction strategies for "hard" and "easy" targets. The server employs sequence-profile alignment (e.g., PSI-BLAST) and profile-profile alignment (e.g., HHSearch) methods to decide whether the query sequence is an "easy" or a "hard" target. Once the nature of the target is identified, MUFOLD-MD uses different methods to generate a sufficiently large set of models and finally to rank them.

For hard targets, models (~8,000) were generated using the Rosetta 3.1 software[1-4] (*ab-initio* method). To this end, secondary structure information from the amino acid sequence was obtained by using PSIPRED[5] and fragment libraries were built from the NCBI database files. The decoys for CASP9 were generated on 47 dual-core Intel Xeon EM64T-2.8GHz CPUs. Only 94 of the lowest Rosetta energy structures were retained, and then further refined by using the "relax" mode in Rosetta 3.1. Finally, the 94 refined full-atom models were ranked by employing the MDR method, and the obtained top 5 structures were submitted to the CASP9 assessors.

For easy targets, around two thousands models were generated by using the Multi-dimensional Scaling (MDS) method[11], and subsequently ranked according to their OPUS_Ca[6] scores. Again, the top 94 structures were retained for further refinement (using Rosetta 3.1) and final ranking using the MDR method.

Our MDR method consists of several important steps. First, an all-atom, high-resolution structure is built for each of the 94 predicted structures. For this, missing H atoms are added by using PSFGEN, which is part of the visual molecular dynamics (VMD) package[7]. Next, the obtained structures are optimized by removing the bad contacts through energy minimization. Finally, the stability of the structures is tested by monitoring the change of their RMSD (with respect to their low-resolution structures) during the MD simulation of their scheduled heating at a rate of 1 K/ps. The MD simulations are carried out in vacuum by coupling the system to a Langevin heat bath whose temperature can be varied according to a desired protocol. All energy minimization and MD simulations were performed by employing the CHARMM force field[8,9] and the parallel NAMD2.6 MD simulation program[10]. Based on extensive testing of the MDR method we have found that statistically the best ranking parameter of the predicted structures is

their mean RMSD during heating from 40K to 140K. This can be achieved through 100ps long MD simulations that take a matter of hours on a single dual core Intel Xeon EM64T-2.8GHz CPU.

The MUFOLD-MD server was used for protein structure prediction in the CASP9 competition. Once the native structures for the CASP9 targets were released we were able to assess the quality of our predicted structures and the efficiency of each part of our MUFOLD-MD server. The results of this analysis will be presented during the CASP9 meeting.

1. Bonneau, R., Strauss, C. E. M., Rohl, C. A., Chivian, D., Bradley, P., Malmström, L., Robertson, T. Baker, D. (2002) *Journal of Molecular Biology* **322,** 65-78.
2. Bonneau, R., Tsai, J., Ruczinski, I., Chivian, D., Rohl, C., Strauss, C. E. M. & Baker, D. (2001) *Proteins: Structure, Function, and Genetics* **45,** 119-126.
3. Simons, K. T., Ingo Ruczinski, Kooperberg, C., Fox, B. A., Bystroff, C. & Baker, D. (1999) *Proteins: Structure, Function, and Genetics* **34,** 82-95.
4. Simons, K. T., Kooperberg, C., Huang, E. & Baker, D. (1997) *Journal of Molecular Biology* **268,** 209-225.
5. Jones, D. T. (1999) *Journal of Molecular Biology* **292,** 195-202.
6. Wu Y, Lu M, Chen M, Li J, Ma J (2007) *Protein Sci 2007* **16(7),** 1449-1463.
7. Humphrey, W., Dalke, A. & Schulten, K. (1996) *J. Mol. Graphics* **14,** 33-38.
8. MacKerell Jr, A. D., Bashford, D., Bellott, M. & others (1992) *FASEB J.* **6,** A143-A143.
9. MacKerell Jr, A. D., Bashford, D., Bellott, M. & others (1998) *J. Phys. Chem. B* **102,** 3586--3616.
10. Phillips, J. C., Braun, R., Wei Wang, Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kalé, L. & Schulten, K. (2005) *Journal of Computational Chemistry* **26,** 1781-1802.
11. Zhang, J., Wang, Q., Barz, B., He, Z., Kosztin, I., Shang, Y. & Xu. D. (2009) *Proteins: Structure, Function, and Bioinformatics* **78**, 1137-1152.

MUFOLD-QA

Removing Redundant Models in Consensus-Based QA

Q. Wang[1], K. Vantasin[1], J. Zhang[1,3], I. Kosztin[2], D. Xu[1,3], and Y. Shang[1]
[1]Department of Computer Science, [2]Department of Physics and Astronomy, [3]Christopher S. Bond Life Sciences Center, University of Missouri,
Columbia, MO 65211, USA
qwp4b@mail.mizzou.edu, shangy@ missouri.edu

In consensus-based quality assessment of protein structures, an important issue is the selection of the reference models that the consensus is built upon. Typically each predicted structure to be evaluated is compared against the reference models and the average similarity of a predicted structure to the references is used as the predicted quality of this structure.

Preferably, the set of reference models mainly consists of near-native structures. Unfortunately, the selection of high-quality models is not a trivial task without the knowledge of native structures. In the past, various approaches have been introduced to determine appropriate reference structures. As an example, MULTICOM[1], one of the top QA servers in CASP8, utilized score functions to pick reference models. However, due to the inaccuracy of existing (energy based or statistics based) score functions and noisy nature of the scores, the performance of MULTICOM, as well as other top CASP8 QA methods, is even inferior to the method using all server predictions as references.

Methods

The inability of existing selection approaches to pick good reference models motivated our investigation into a new approach, which is based on an observation that protein tertiary structure prediction usually generates redundant models, often produced artificially due to using the same software or method. Using structures with multiple duplicate copies as references may make the reference biased towards these structures in the consensus method. This is especially true for CASP data because a participating team may register multiple servers and for a target each server is allowed to submit up to five models, among which some may be identical or highly similar in an artificial way for some targets. Hence, to improve the quality of a reference set, one way is to deprive its redundancy. This is the main idea behind our reference model selection approach in MUFOLD-QA.

Specifically, to extract reference models from a set of predicted structures of a target, we compare the pairwise GDT-TS score between each pair of predicted structures s_i and s_j to a threshold Z, which was learned from previous CASP data. If the pairwise GDT-TS score is greater than Z, we consider either s_i or s_j redundant and randomly remove one of them from being a reference. By checking each pair of predicted models and discarding the redundant one, a reference set is constructed.

Based on this method, we developed a fully automatic server MUFOLD-QA that predicts global quality of CASP9 models using a consensus approach. MUFOLD-QA works as follows:

(1) Download the server predicted structures of a CASP9 target.
(2) Determine the value of the threshold Z for this target based on the pool of predicted structures.
(3) Extract reference models from the set of predicted structures by discarding redundant ones.
(4) Compute GDT-TS scores between each predicted structure to each of the reference models. The predicted quality of a structure is its average GDT-TS score to references.
(5) Submit the predicted quality of all server structures.

In addition to randomly discarding one of a pair of similar structures in the reference selection process, we also tried some deterministic methods, such as removing all pairs with GDT-TS scores greater than *Z*. Their results were not as good empirically.

Another way of removing redundancy is to assign different weights to reference models so that the models with more duplicates receive smaller weights. The weighted sum of the pairwise GDT-TS score of a structure to references is the predicted quality of the structure. We tried different ways to assign weights and the results were slightly worse than the method implemented in MUFOLD-QA.

Results

Using CASP7 data as the training set, we determined the parameters of MUFOLD-QA. Using CASP8 data as the test set, we compared MUFOLD-QA with the best QA servers in CASP8, the best scoring functions we can get, and the consensus method using all server predictions as references. The results showed that the QA scores of MUFOLD-QA correlated better with the true GDT-TS scores than any of the other methods.

Additionally, a preliminary evaluation of CASP9 QA results (http://sysbio.rnet.missouri.edu/casp9_assess/qa.php) shows that MUFOLD-QA is one of the top QA servers in CASP9.

1. J. Cheng, Z. Wang, A. N. Tegge and J. Eickholt. Prediction of global and local quality of CASP8 models by MULTICOM series. *Proteins: Structure, Function, and Bioinformatics*, 77(S9), pp. 181-184, 2009.

MUFOLD-WQA

Taking the Middle Path: A Band-Pass Consensus QA Method

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

As demonstrated in previous CASP experiments, consensus QA methods were very effective, outperforming single-model methods by a large margin. In the consensus QA approach, the predicted quality score of a model is typically the average of some similarity measurement of the model to a set of reference models. For CASP, a basic consensus QA method, as called ***total consensus QA***, is to use all server models as the reference set. In CASP8, the differences among the top QA servers were mostly in the reference sets they selected. Surprisingly, the simple total consensus QA performs better than all top CASP8 QA servers in our own test using the CASP8 data. Apparently, although the teams spent significant time looking for good references, it not easy to find a strategy that beats the one simply using all server models as references.

Through systematic and in-depth analysis of different reference selection strategies, we discovered two important insights: 1) use different reference sets for different models and 2) weight different reference models differently and, in particular, do not use models that are very similar or too different from the candidate model pool as references. We experimented with a series of reference selection functions and ended up using a simple, yet effective band selection function in our MUFOLD-WQA server. Preliminary CASP9 result shows that MUFOLD-WQA performed very well and is tied with two other servers at the top in the QA category. In addition, MUFOLD-WQA is the best server for model selection.

Methods

Given server predicted models for each target, MUFOLD-WQA computes a quality score for each model based on its similarity to other models. The similarity measurement can be any commonly used metric, such as GDT-TS, TM-score, or Q-score. We tested these metrics on the CASP8 data set and found that GDT-TS gave slightly better result. For each model, after computing its similarity to all other server models, we either perform a weighted averaging of the similarity measurements, e.g., GDT-TS, or apply a selection function to choose a subset of models and do a simple averaging of the similarity measurements to them.

One weight function we tried is the ***sigmoid*** function, $sig(x) = \frac{1}{1+e^{c(x-\alpha)}}$, where x in the range [0, 1] is the similarity of the candidate model to another server model, and c and α are constants. When c is large, the function becomes a simple ***step*** selection function, $step(x) = 0$, if $x \geq \alpha$; 1, if $x < \alpha$, where $0 < \alpha < 1$. Generalizing the step function, we have the ***band*** selection function:

$band(x) = 1$, if $a < x < b$; 0, otherwise.

Parameters of these functions were determined empirically using the CASP8 data. For different types of targets, whether they were easy or hard to predict by the servers, the best parameters were different. We divided the targets into three classes, easy, medium, and hard, based on some indicators, such as the average pair-wise GDT-TS score of all server models. Through systematic experiments using the CASP8 data, we determined the best parameters for targets in the three classes and fixed them in the CASP9 runs. We found that although the sigmoid function is flexible, the step function with good parameters can be as good as the sigmoid function. Still, the band function is better than the step function. We liked the simplicity of the band function and implemented it in our MUFOLD-WQA server.

In addition to QA as CASP defined, our method also performs model selection based on the generated quality score well.

Results

MUFOLD-WQA significantly outperformed all top QA methods and the total consensus QA on the CASP8 data. For CASP9, the table below shows a preliminary QA results based on 93 known targets [1]. MUFOLD-WQA is tied with two other servers as #1 in the QA category with average correlation between our predicted scores and the true GDT-TS values of 0.92.

Rank	Predictor	Average Corr.	Num of Targets
1	QMEANclust	0.92	93
1	MUFOLD-QA	0.92	93
1	**MUFOLD-WQA**	**0.92**	**93**
4	MULTICOM-cluster	0.915	93
5	MetaMQAPclust	0.896	93

For model selection, using CASP8 data, we compared MUFOLD-WQA with existing state-of-the-art scoring functions, including OPUS-Ca, ModelEvaluator, DFIRE, RAPDF and DOPE, all top QA servers in CASP8, and the total consensus QA method. On average, the top 1 ranked models selected by MUFOLD-WQA are better than those ranked by any other methods. In CASP9, the table below shows the best model selection results -- the sum of GDT-TS of the top models selected by various QA servers based on 93 known targets [1]. Again, MUFOLD-WQA is the best.

Rank	Predictor	Sum of GDT-TS	Num of Targets
1	**MUFOLD-WQA**	**54.939**	**93**
2	MULTICOM-cluster	54.8999	93
3	QMEANclust	54.7996	93
4	IntFOLD-QA	54.7582	93
5	ModFOLDclust2	54.723	93

1. http://sysbio.rnet.missouri.edu/casp9_assess

MULTICOM
MULTICOM-CLUSTER
MULTICOM-CONSTRUCT
MULTICOM-NOVEL
MULTICOM-REFINE

Sequence and model-based prediction of protein residue-residue contacts by MULTICOM predictors

Jesse Eickholt[1], Zheng Wang[1], and Jianlin Cheng[1, 2, 3]
[1] – Computer Science Department , [2] – Informatics Institute,[3] – C. Bond Life Science Center, University of Missouri, Columbia, MO 65211 USA
chengji@missouri.edu

We present our MULITCOM series of protein residue contact predictors. They span the full spectrum of contact prediction approaches, including sequence-based machine learning, contact-map post processing and model-based consensus methods.

Methods

MULTICOM-NOVEL and MULTICOM-REFINE are sequence-based, *ab-initio* methods based on our recursive neural network predictor, NNcon[1]. The basis of this software package is a set of recursive neural network ensembles, one which predicts general residue-residue contacts and another trained specifically to predict beta-residue pairings in beta-sheets. Features used for each residue include a sequence profile, secondary structure and solvent accessibility. MULTICOM-NOVEL used only general residue-residue contact predictions, whereas MULTICOM-REFINE combined specific beta-residue contact predictions with general residue contact predictions.

MULTICOM-CLUSTER is a sequence-based, *ab-initio* approach based on our SVM predictor, SVMcon[2, 3]. Here, for each residue-residue pair in the target sequence, a set of features including secondary structure, solvent accessibility and a sequence profile is encoded for a 9-residue window centered around each residue. This feature vector is fed into a support vector machine (SVM) trained on a large dataset which classifies the residue-residue pair as "in contact" or "non-contact". Those pairs classified as "in-contact" were submitted as predictions.

MULTICOM-CONSTRUCT started with a contact map generated by MULTICOM-NOVEL. A contact map for a target n residues in length is an n x n matrix where the i,j^{th} entry is a value representing the predicted probability of residues i and j being in contact. Using the contact map, we figured a score for each possible residue-residue pair. More specifically, to calculate the contact score for residues i and j, the Pearson correlation coefficient for the i^{th} and j^{th} rows of the contact map was calculated and scaled to be from 0 to 1. This was done based on our assumption that two residues in contact have similar spatial contact relationships with other residues. The scaled value was used to rank all residue-residue contacts and the top scoring pairs were submitted as predictions.

Our human predictor MULTICOM uses an automated, model-based consensus approach to make predictions. Extracting contacts from tertiary structure predictions is in and of itself nothing new as past CASP evaluations have inferred contacts from models and ranked them according to C_β-C_β distance, or Cα for glycine[4]. This process, however, is very limited in that it only considers one model at a time. Consequently, it is unable to utilize the complementary information contained in multiple models generated from varying methods and groups. Our MULTICOM predictor is based on a novel consensus voting approach which extracted contacts from all tertiary structure models submitted for a target and counted the number of times a residue-residue pair was in contact across the various models. These

contact counts were scaled, ranked and then submitted as the predicted residue-residue contacts. Surprisingly, this simple approach works extremely well according to our preliminary assessment.

Results

We performed a preliminary evaluation of our predictors on 10 targets which were available at the time of writing and are putative *ab-initio* targets according to our analysis (i.e., no significant templates could be found for these targets). This preliminary evaluation is based on the precision of predicted contacts at two separation thresholds. The precision is defined as the percentage of correct predictions. When calculating the precision, we first ranked all submitted residue-residue predictions for a target by score and then considered only the top L/5 (i.e., total target length divided by 5) predictions for each target.

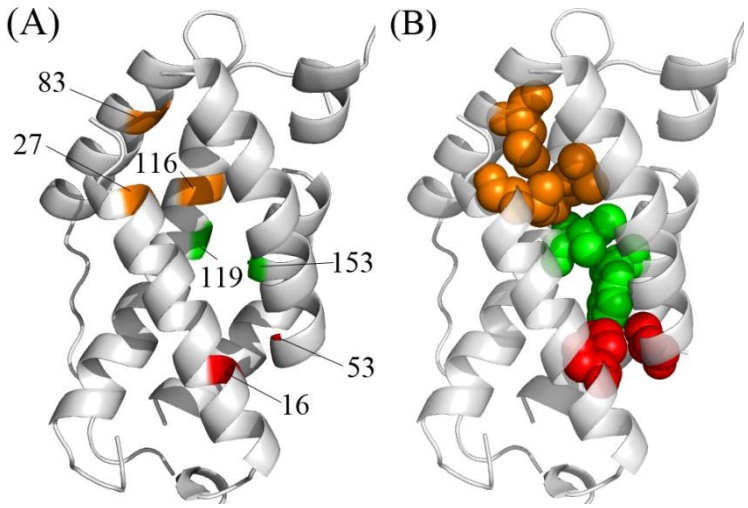
Table 1 reports the preliminary assessment of medium- and long-range contact predictions of our contact predictors. The median precision for all CASP9 contact predictors in the server category was .16 at both thresholds, indicating that our full range of contact predictors participated competitively in this round of CASP. Our human contact predictor, MULTICOM, showed particular promise as it outperformed all server predictors at a variety of residue separation thresholds (detailed data not shown). For instance, MULTICOM has relatively high precision (i.e., 39%) for long-range contacts with sequence separation >= 24, which may be useful for constructing and evaluating models for hard *ab-initio* targets. However, whether or not these results demonstrate that CASP9 server predictors as whole would be able to generate accurate contact restraints sufficient for reconstructing tertiary structures for *ab initio* targets has yet to be proven. **Figure 1** visualizes several long-range contacts of a hard target T0618 correctly predicted by MULTICOM.

A more in-depth listing of our preliminary results for all CASP9 residue contact predictors in the server category can be found at http://sysbio.rnet.missouri.edu/casp9_assess/contact.html .

Table 1. Preliminary results of MULTICOM series residue contact predictors. We evaluated our protein residue contact predictors on 10 targets which we believe to be *ab-initio* targets (T0531, T0534, T0550, T0555, T0578, T0581, T0618, T0621, T0624, and T0637). Separation thresholds are in number of residues. Precision reported is for top L/5 predicted contacts.

Predictors	Precision at Sep. >=12 and < 24	Precision at Sep. >= 24
MULTICOM	.48	.39
MULTICOM-CLUSTER	.28	.24
MULTICOM-REFINE	.20	.17
MULTICOM-CONSTRUCT	.21	.16
MULTICOM-NOVEL	.17	.13

Figure 1. Examples of long-range residue contacts for T0618 as predicted by MULTICOM. (A) Some key long range contact predictions for target T0618. The following pairs of residues were predicted in contact: residue 16-53 (red), 119-153 (green), 27-116 (orange), and 83-116 (orange). (B) illustrates the spheres of the contact residues after depicting all the atoms.



Availability

The MULTICOM-CLUSTER software and web service (i.e., SVMcon server) are available at http://casp.rnet.missouri.edu/svmcon.html. The MULTICOM-REFINE/NOVEL software and web service (i.e., NNcon server) are available at http://casp.rnet.missouri.edu/nncon.html.

1. Tegge, N. Wang, Z., Eickholt, J. & Cheng, J. (2009). NNcon: Improved protein contact map prediction using 2D-Recursive neural networks. *Nucleic Acids Research.* **37,** w515-w518.

2. Cheng, J. & Baldi, P. (2007). Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinformatics*. **8**,113.

3. Vapnik, V.N. (1995). *The Nature of Statistical Learning Theory*. New York: Springer-Verlag.

4. Ezkurdia, I., Graña, O., Izarzugaza, J. M. G. & Tress, M. L. (2009). Assessment of domain boundary predictions and the prediction of intramolecular contacts in CASP8. *Proteins.* **S9**, 196-209.

MULTICOM-CLUSTER
MULTICOM-REFINE
MULTICOM-NOVEL
MULTICOM-CONSTRUCT
MULTICOM

Integrated Prediction of Protein Tertiary Structure by MULTICOM Predictors

Jianlin Cheng, Zheng Wang, and Jesse Eickholt
Department of Computer Science, University of Missouri, Columbia, MO, USA
chengji@missouri.edu

Our group tested four automated servers (i.e., MULTICOM-CLUSTER, MULTICOM-REFINE, MULTICOM-NOVEL, and MULTICOM-CONSTRUCT) and one semi-automated human predictor (i.e., MULTICOM) for tertiary structure prediction. Our servers integrated multiple templates, alignments, and models to generate models and combined template-based and template-free modeling to refine models. Our human predictor focused on model refinement driven by model quality assessment. Here we briefly describe these methods with an emphasis on new developments since CASP8.

Methods

Our automated protein structure prediction pipeline consists of the following four steps: (1) template-ranking and alignment, (2) template combination, (3) model generation and assessment, and (4) model combination and refinement. The server predictors differed in the last two steps. The human predictor focused on model assessment and refinement.

Template ranking and alignment. The servers used several standard sequence and profile alignment methods, including BLAST, PSI-BLAST, HMMer, SAM, HHSearch, Compass, and PRC, to identify templates and generate query-template alignments. Each alignment method searched a query sequence against the template sequence and profile databases for significant template hits. The top 10 template hits ranked by the e-values of the query-template alignments were retained for each method and the query-template alignments from the top hits identified by each method were stored in separate lists for further analysis. Furthermore, the servers counted the number of times a template was found by each alignment method and generated a consensus list of the top 10 templates ranked solely by the frequency counts.

Multiple template combination. The MULTICOM servers used a new *structure-alignment-guided, central-star, top-down approach* to combine each list of query-template alignments. The method first selected a top ranked query-template as a seed. Using the common query sequence as an anchor, it combined other template-query alignments ranked lower in the list with the seed if their e-values were close to the seed alignment and their aligned regions were structurally consistent with previously combined query-template alignments. The structural similarity of two query-template alignments was checked by comparing the structure of two templates which align to the same regions of the query (as determined by TM-Align[1]). The *structural consistency check* is a new addition to the alignment combination method[2] used in CASP8. It can ensure the structural consistency of combined templates which improves model quality by avoiding or reducing atom clashes that result from the combination of structurally inconsistent templates.

For the consensus list of templates, MULTICOM servers used another new *structure-alignment-driven profile-alignment* method to generate structurally consistent alignments between a query and multiple templates. For each template in the list, the method first aligned its structure with that of each of

the remaining templates using TM-Align. Each pairwise template-template structure alignment was converted into a pairwise sequence alignment by retaining only structurally aligned residues in the template. These pairwise sequence alignments between the common template and other templates in the list were combined into a multiple sequence alignment using the common template as an anchor. Because only those regions of the other templates which aligned well to the anchor template were kept, the multiple sequence alignment is *structurally* consistent. The multiple sequence alignment of these templates was then aligned with the profile (or multiple sequence alignment) of the query to generate an alignment between the query and all the templates using the multiple sequence alignment tool MUSCLE. Finally, a list of combined query-template alignments was generated for the consensus template list.

Model generation and assessment. Each combined query-template alignment and the associated template structures were fed into Modeller[3] to simulate 10 conformations and the one with the lowest energy (as calculated by Modeller) was used as a model. All the models were pooled together to be assessed by four different strategies.

MULTICOM-CLUSTER used an *ab initio* model evaluation method *ModelEvaluator*[4] to predict the GDT-TS score of each model and ranked them accordingly. Each model in the pool was then compared with the top five ranked models. The average GDT-TS score between the model and the top five models was the final predicted GDT-TS score of the model[5]. These scores were used to re-rank the models. MULTICOM-REFINE used a pairwise model comparison method to rank models. It compared each model against all other models using TM-score[6]. The average of the max-sub scores between the model and other models was used as the predicted quality of the model and the predicted quality scores were used to rank models. MULTICOM-NOVEL used a new *multi-level model selection approach* to select models. At level one, it chose models by sequence identity between the query and the top templates if identity was more than 0.4. At level two, it chose models based on the frequency of the top templates appearing in the template lists if the maximum template frequency was more than 3. At level three, it chose models based the scores of pairwise model comparison if the maximum pairwise score was more than 0.4. Finally, at level four, it selected models based on *ab initio* model quality scores predicted by ModelEvaluator. MULTICOM-CONSTRUCT selected only the top two models generated from each template list based on the e-values of top query-template alignments. These selected models were then ranked by their pairwise quality scores. Our MULTICOM human predictor used both our own server models and the models produced by other CASP9 server predictors as input. It also collected the quality scores of these models predicted by dozens of CASP9 quality assessment (QA) predictors. The averages of the predicted quality scores were used to rank these models first. Then each model was compared with the top five ranked models to generate an average structure comparison score as its predicted quality score[5]. The top five models were selected for refinement.

Model combination and refinement. The pipeline used two methods to refine our server models. Firstly, it used a top-down local-global model combination approach[7] in MULTICOM-REFINE to combine the top ranked models with other models that were globally very similar to it (e.g., pairwise GDT-TS score > 0.7) or only the very similar local regions of other models if no globally similar models were found. Secondly, in order to improve the quality of unfolded tail regions of some models, MULTICOM servers used a new, *hybrid* method to integrate template-free modeling and template-based modeling to refine models. For the models generated from template alignments not covering the long tails (e.g., >= 15 residues) of a query sequence, MULTICOM servers used a modified fragment-assembly method[8] to rebuild the peripheral tails. This method took the internal core region modeled by template-based modeling into consideration when calculating the energy but kept the core rigid. *This approach can integrate template-based and template-free modeling at an arbitrary percentage*. Our semi-automated human predictor only used the top-down global-local model combination to refine models since no alignment information was available.

Results

We assessed our four server predictors, our human predictor and other CASP9 server predictors on the entire chain of 93 targets whose experimental structures were released to date. We used TM-score to compare CASP9 models against the target structures and to calculate the GDT-TS scores and TM-scores after the removal of disordered regions. Table 1 reports the results of the MULTICOM predictors. A preliminary assessment of all CASP9 server predictors is available at: http://sysbio.rnet.missouri.edu/casp9_assess/ . The results on these targets seem to show that our top server predictor is ranked among the top CASP9 server predictors.

Table 1. The average GDT-TS and TM scores of top one and best of five models on 93 targets.

Predictors	Top One			Best of Five	
	GDT-TS	TM-Score		GDT-TS	TM-Score
MULTICOM (human)	0.587	0.654		0.600	0.667
MULTICOM-CLUSTER	0.552	0.619		0.575	0.642
MULTICOM-NOVEL	0.548	0.613		0.564	0.629
MULTICOM-REFINE	0.549	0.613		0.567	0.631
MULTICOM-CONSTRUCT	0.547	0.611		0.568	0.633

Availability

The servers are at http://casp.rnet.missouri.edu/multicom_3d.html.

1. Zhang, Y., Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on TM-score. *Nucleic Acids Research*. **33**, 2302-2309.
2. Cheng, J. (2008). A multi-template combination algorithm for protein comparative modeling. *BMC Structural Biology*. **8**, 18.
3. Sali, A., Blundell, T. Comparative protein modeling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779-815.
4. Wang, Z. et al.. (2009). Evaluating the absolute quality of a single protein model using support vector machines and structural features. *Proteins*. **75**, 638-647.
5. Cheng, J. et al.. (2009). Prediction of global and local quality of CASP8 models by MULTICOM series. *Proteins*. **77**, 181-184.
6. Zhang, Y., Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**, 702-710.
7. Wang, Z. et al.. MULTICOM: a multi-level combination approach to protein structure prediction and its assessment in CASP8. *Bioinformatics*. **26**, 882-888.
8. Simons, K.T. et al.. (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.

MULTICOM-NOVEL
MULTICOM-CLUSTER
MULTICOM-REFINE
MULTICOM-CONSTRUCT
MULTICOM

Evaluating the quality of single and multiple protein models by MULTICOM quality assessment predictors

Zheng Wang[1], Jesse Eickholt[1], and Jianlin Cheng[1, 2, 3]
[1] – Computer Science Department , [2] – Informatics Institute,[3] – C. Bond Life Science Center, University of Missouri, Columbia, MO 65211 USA chengji@missouri.edu

We develop and test five model quality assessment (QA) predictors: MULTICOM-NOVEL, MULTICOM-CLUSTER, MULTICOM-REFINE, MULTICOM-CONSTRUCT, and MULTICOM. MULTICOM-NOVEL evaluates the absolute quality of a single model in terms of GDT-TS score. MULTICOM-CLUSTER assesses the quality of a group of models by pair-wise model comparisons of a pool of models. MULTICOM-REFINE utilizes a hybrid approach to refine the quality scores initially generated by MULTICOM-NOVEL. MULTICOM-CONSTRUCT performs a similar refinement process but on average pair-wise model similarity scores as measured by Q-scores. Our human predictor MULTICOM makes consensus evaluations by refining the average quality scores predicted by all CASP9 QA predictors.

Methods

MULTICOM-NOVEL is based on our *ab initio* QA predictor - ModelEvaluator[1]. Given a single model, ModelEvaluator extracts secondary structure, solvent accessibility, beta-sheet topology, and a contact map from the model and then compares these items with those predicted from the primary sequence using the SCRATCH program[2]. These comparisons generate match scores which are then fed into a SVM model trained on CASP6 and CASP7 data to predict the absolute quality of the model in terms of GDT-TS scores. In order to avoid mistakenly ranking highly scored *ab initio* models ahead of template-based models, for each target, the top two models ranked by ModelEvaluator are compared with that of MULTICOM-CLUSTER (see the following paragraph for details). If the average GDT-TS score between them is less than 0.6, the output of ModelEvaluator is replaced by that of MULTICOM-CLUSTER. During CASP9, the replacement was performed only on T0533 and T0559.

MULTICOM-CLUSTER is a pair-wise model comparison approach[3]. Taking a pool of models as input, it first filtered out illegal characters and chain-break characters in their PDB files. It then used TM-SCORE[4] to perform a full pair-wise comparison between these models. The average GDT-TS score between a model and all other models is used as the predicted GDT-TS score of the model. One caveat is that the GDT-TS score of a partial model is scaled down by the ratio of its length divided by the full target length.

MULTICOM-REFINE used a hybrid approach[5] to integrate *ab inito* model ranking methods with structural comparison-based methods. It first selects several top models (i.e. top five or top ten) as reference models. Each model in the ranking list is superposed with the reference models by TM-SCORE. The average GDT-TS score of these superimpositions is considered as the predicted quality score. The superimpositions with the reference models are also used to calculate Euclidean distances between the same residues in the superimposed models. The average distance is used as the predicted local quality of the residue.

MULTICOM-CONSTRUCT first uses the average pair-wise similarity scores, calculated in terms of Q-score[6, 7], to generate an initial ranking of all the models. The Q-score between a pair of residues (*i, j*) in two models is computed as:

$$Q_{ij} = \exp[-(r_{ij}^a - r_{ij}^b)^2]$$

where r_{ij}^a and r_{ij}^b are the distance between C_α atoms at residue position *i* and *j* in model *a* and *b*, respectively. The overall Q-score between model *a* and *b* is equal to the average of all Q_{ij} scores of all residue pairs in the entire model. The average Q-score between a model and all other models is used as the predicted quality score of the model. The initial quality scores are refined by the same refinement process used by MULTICOM-REFINE.

MULTICOM is a consensus approach. It downloads all the predictions made by CASP9 QA predictors. The QA scores are averaged as consensus quality scores for these models, which are used to generate an initial ranking of the models. The same refinement process used by MULTICOM-REFINE is applied to refine these scores to generate both local and global quality predictions.

Results

We preliminarily assess the CASP9 QA predictors on the experimental structures of 93 targets released by the time of writing this abstract. We download all the CASP9 tertiary structure (TS) models from the CASP9 web site and the experimental structures from the Protein Data Bank (PDB). These PDB files are preprocessed in order to select correct chains and residues that match the CASP9 target sequences. TM-SCORE is used to align each TS model with the corresponding native structure to generate its real GDT-TS score which is treated as the actual model quality scores. We download all the CASP9 QA predictions and evaluate them against the actual quality scores by four criteria: average per-target correlation[8], the average sum of the GDT-TS scores of the top one ranked models, the overall correlation on all targets[8], and the average loss[5, 8] (**Table 1**). The loss is defined as the average difference between the GDT-TS score of the overall best model and the GDT-TS score of the top model ranked by a QA predictor. Both the average sum of the GDT-TS scores of the top one ranked models and the loss can assess the ranking abilities of a QA predictor, i.e. whether it can rank high-quality models at the top.

Table 1. The average per-target correlation, the average sum of GDT-TS scores of top one ranked models, the overall correlation, and the average loss of our methods on 93 CASP9 targets.

Predictors	Avg. Corr.	Avg. Top 1	Over. Corr.	Avg. Loss
M.-NOVEL	0.680	0.55	0.763	0.092
M.-CLUSTER	0.915	0.59	0.942	0.057
M.-REFINE	0.867	0.56	0.926	0.083
M.-CONSTRUCT	0.832	0.57	0.900	0.076
MULTICOM	0.882	0.57	0.924	0.060

Besides our five QA predictors, we also evaluate all the other CASP9 QA predictors. The preliminary assessment shows that MUTLICOM-CLUSTER is one of top CASP9 QA predictors in terms of the four criteria. The preliminary evaluation results can be accessed at: http://sysbio.rnet.missouri.edu/casp9_assess/ .

Availability

We plan to release an executable of MULTICOM-CLUSTER and -CONSTRUCT at http://casp.rnet.missouri.edu/~chengji/cheng_software.html in the near future.

1. Wang, Z., Tegge, A. & Cheng, J. (2008). Evaluating the absolute quality of a single protein model using structural features and support vector machines. *Proteins* **75**, 638-647.
2. Cheng, J., Randall, A., Sweredoski, M. & Baldi, P. (2005). SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Research* **33**, W72-W76.
3. Larsson, P., Skwark, M.J., Wallner, B. and Elofsson, A. (2009). Assessment of global and local model quality in CASP8 using Pcons and ProQ. *Proteins* **77**, 167-172.
4. Zhang, Y. & Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702-710.
5. J. Cheng, Z. Wang, Tegge, A. & J. Eickholt (2009). Prediction of Global and Local Quality of CASP8 Models by MULTICOM series**.** *Proteins* **77**, 181-184.
6. Ben-David, M., Noivirt-Birk, O., Paz, A., Prilusky, J., Sussman, J., Levy, Y. & Pearl, E. (2009). Assessment of CASP8 structure predictions for template free targets. *Proteins* **77**, 000-000.
7. McGuffin, L. & Roche, D. (2010). Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments. *Bioinformatics* **26**, 182-188.
8. Cozzetto, D., Kryshtafovych, A., Tramontano, A. (2009). Evaluation of CASP8 model quality predictions. *Proteins* 77, 157-166.

MUSICS_server

Multigrid Sequential Importance Sampling: A New Sampling Method for Protein Simulations

K Bartz[1], D Fernandez[1], S Wong[1], P Edlefsen[1], S C Kou[1], J S Liu[1] and J Zhang[2]
[1]*Department of Statistics, Harvard University,* [2]*Department of Statistics, Florida State University*
jinfeng@stat.fsu.edu, kou@stat.harvard.edu, and jliu@stat.harvard.edu

Sampling is widely seen as the current bottleneck in protein structure prediction. In this study, we developed a new sampling method, MUltigrid Sequential ImportanCe Sampling (MUSICS), based on a Monte Carlo method we designed previously, called FRESS (fragment re-growth via energy-guided sequential sampling). FRESS[1] has been shown to perform very well on HP model, a relatively simple model, but still quite challenging in terms of finding conformations with global minimum energy. MUSICS is an implementation of FRESS algorithm on atom-level realistic protein models. We tested MUSICS on the CASP9 experiment and obtained encouraging results. We also compared MUSICS with CCD (Cyclic Coordinate Descent) in terms of the efficiency in loop closure.

Methods

The core of our sampling method is an efficient way to sample the conformation of a loop (or a fragment) of a protein with the conformation of the rest of protein chain fixed, also called a fixed-ends move. It is an essential step for most Monte Carlo simulation of protein structures and has been studied by many researchers in the past. The state-of-the-art method is CCD (Cyclic Coordinate Descent) developed by Canutescu and Dunbrack[2]. The basic idea comes from two previous works, configurational-bias Monte Carlo[3] and Multigrid Monte Carlo[4].

In MUSICS, for a fragment of length L, we regrow the fragment one residue at a time. At each step, we sample a number of possible conformations for residue k. The sampling procedure regrows one residue at a time until reaching the second-to-last residue. The final two residues are placed using an analytic closure algorithm by Coutsias et al[5]. The resulting new conformation of the fragment will then be accepted or rejected using a criterion similar to that used Metropolis-Hastings algorithm. There are two components to the weighting used to sample each residue k. To enourage eventual fragment closure, the first weighting factor is based on empirical end-to-end distances over segments of $L - k$ residues, as observed in a range of proteins in Protein Data Bank. The second weighting factor is the incremental energy impact on the fixed part of the protein chain – that is, every atom except those in the fragment that have not yet been regrown. Each regrown fragment is a proposal in our Markov Chain Monte Carlo (MCMC) simulation of protein structures. We also implemented parallel tempering (replica exchange Monte Carlo) at the top level.

Results

Our sampling method runs rapidly, averaging 5.9 ms to regrow a fragment of length 8 on a 1.6 GHz machine. However, only about 23% of regrown fragments successfully close, which drops the rate to an average 30.7 ms for successfully closed fragments. This is comparable to CCD's reported rate of 37 ms, though in practice we were not able to replicate CCD's speed; our local CCD implementation ran at 256 ms per fragment.

We conducted a comparative experiment using 100,000 random segments of varying length and position from CASP8 proteins. We applied both methods to each segment, running ours 500 times and CCD 50 times on each. Since CCD can be applied to any sampling method, to ensure comparability we used a sampler for CCD that is identical to ours in most respects. The only difference is that it does not incorporate our reweighting of residue selection probabilities by end-to-end distances, which is our

sampler's primary feature to encourage closure. Also note that CCD's closure condition – defined as an RMS under 0.08 for the first three atoms beyond the end of the fragment – is not interpretable in our method because ours does not attempt to modify the subsequent three atoms. Consequently, for comparison we have defined a single set of closure criteria based on six geometric conditions at the boundary.

The leftmost panel of Figure 1 shows the average time each method needed to regrow a single fragment, closed or not. The middle panel shows the average time needed to generate a single closed fragment. Note that not quite all fragments "closed" by CCD are closed according to our criteria, which explains why the blue CCD lines are not identical in the leftmost and middle panels. The rightmost panel shows the average time needed to generate a regrown fragment within an energy threshold of the original, pre-regrown conformation. Any threshold gives similar results; here, we use 5 units of a simple Van der Waals potential. As the figure shows, our method shines most in producing regrown fragments with low energy, outpacing the CCD-based fragments method by a computational factor ranging from 14 for short fragments to about 4 for long fragments. Applied to a folding task, this efficiency helps speed the search for conformations of lower energy.

In our own CASP9 folding entry, we suffered at the start from a poor template detection system and energy function. Our performance improved as we began experimenting with several available template detection methods and developing more accurate energy functions. Although the official CASP9 rankings are not yet available, we used the Zhang lab's public CASP9 assessments to chart our progress over time. Figure 2 shows the rank of our team's first model among the 65 regular CASP9 participants. We were ranked near the bottom until T0550, when we rolled out a new energy function and ramped up our computational resources. We then began to rank at about the middle of the pack until the end of the competition.

Figure 1: Average time to generate regrown fragments by our sampler and CCD, as determined in a simulation experiment spanning 100,000 randomly chosen segments in CASP8 proteins.

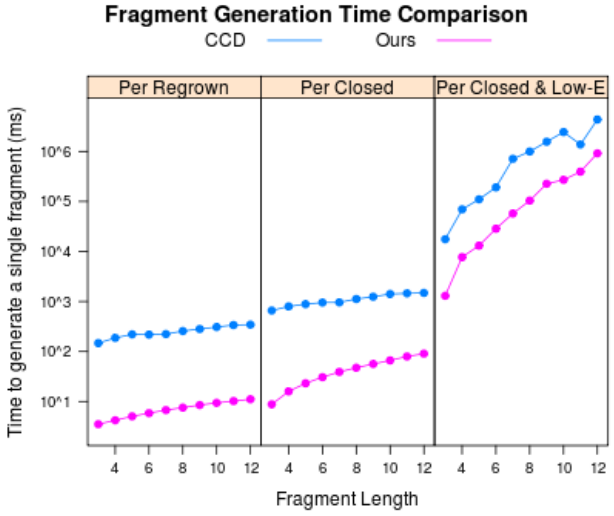
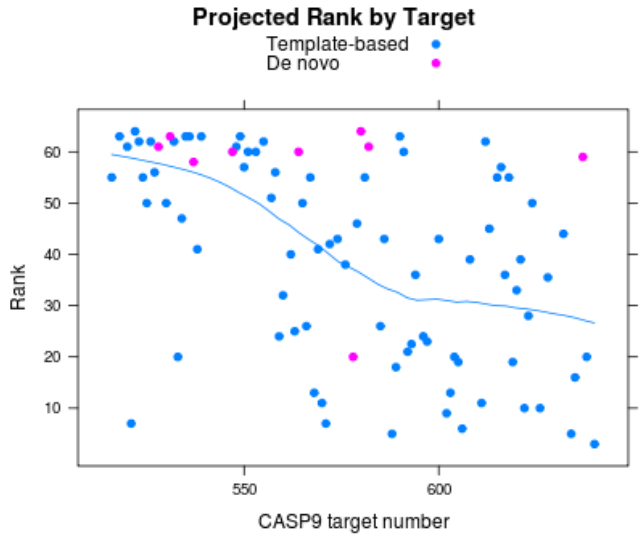


Figure 2: Rank of our team's CASP9 submissions among all teams', as determined independently by Zhang's lab.



Availability

We have implemented the method in C++. The executable is available upon request.

1. Zhang, J., Kou, S.C., Liu, J.S. (2007). Biopolymer structure simulation and optimization via fragment re-growth Monte Carlo. *J Chem Phys*, 126, 225101.
2. Canutescu, A.A., Dunbrack, R. L. (2003) Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Sci.* ***12***, 963-972.
3. Siepmann, J. I. and Frenkel, D., (1992) Configurational-bias Monte Carlo - A new sampling scheme for flexible chains *Mol. Phys*. 75, 59-70.
4. Goodman, J. and Sokal, A. D., (1986) Multigrid Monte Carlo Method for Lattice Field Theories, *Phys. Rev. Lett*. 56, 1015-1018.
5. Coutsias, E. A., Seok, C., Jacobson, M. P., and Dill, K. A., (2004) A kinematic view of loop closure, *J of Comp Chem*, 25:4, 510 - 528.

MUSICS-2S

Loop Modeling using Distance-guided Sequential Monte Carlo

Ke Tang[1], Jinfeng Zhang[2] and Jie Liang[1]
[1] Department of Bioengineering, University of Illinois at Chicago, Chicago, IL, USA, [2]2Department of Statistics, Florida State University, Tallahassee, FL, USA
ktang6@uic.edu

Loop modeling is an important step in protein structure prediction. We developed a new loop sampling method and a loop-specific energy function and applied them in CASP9 on template-based targets.

Methods

Overall strategy. In our method, we first use HHsearch to find homologous proteins to a target sequence. If template can be identified in this step, we then use Modeller[1] to build the initial models. The initial models are then subject to a loop refinement procedure using the loop sampling method and a loop-specific energy function. The final model is selected using a general energy function for whole protein structures.

Loop modeling algorithm. The loop regions in the initial models are first identified. To refine loop regions, we take a new sequential Monte Carlo based sampling approach, called Distance-guided Sequential Monte Carlo (DSMC) for generating loop conformations with lower energy. In this method, we resample the conformations of loops one residue at a time. For each residue, multiple trial conformations are sampled, from which one conformation is selected according to both the energy and how likely the loop will successfully connect. To evaluate the likelihood of a successful re-connection, we use distance propensities calculated from native protein structures in PDB. The conformations of the last two residues are calculated using the CSJD algorithm of Coutsias[2]. Instead of refining loops one by one, we refine all the loops together by an iterative procedure.

Loop-specific energy function. We used an atom level distance dependent knowledge-based energy function. To derive a loop-specific energy function, we obtained a large set of loop conformations and adopted a decoy-based reference state method to obtain the parameters[3]

Results

Our loop modeling approach works well for the Sali's test datasets. The average global backbone root-mean-square deviations (RMSDs) to the native structures are 0.38 A for 5 residue loops and 1.24 A for 9 residue loops for 20 proteins each. In addition, our method can generate long loops with small RMSDs to the native loop conformation (eg. <3A for loops of length 20).In CASP9, our group ranked 66th in 81 groups with one target ranked 1st in 111 targets using Yang Zhang's TM score. We missed more than ten targets at the beginning. The total number of submitted targets is 89 out of 111.

Availability

Not available for public so far.

1. A. Fiser, R.K. Do, & A. Sali. Modeling of loops in protein structures, Protein Science 9. 1753-1773, 2000.
2. EA.Coutsias, C.Seok, MP.Jacobson and KA.Dill. A kinematic View of loop closure. 2004. J Comput Chem. 25:510-528.
3. GY.Chuang, D.Kozakov, R.Brenke, SR.Comeau, S.Vajda. DARS (Decoys As the Reference State) potentials for protein-protein docking. 2008. Biophys J. 1;95(9):4217-27.

MUSTER

MUSTER: a single threading program using sequence and structure profile-profile alignments

Sitao Wu[1] and Yang Zhang[2]
[1]-Center for Research in Biological Systems, University of California, San Diego, 9500 Gilman Dr., La Jolla, CA 92093, [2]- Center for Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw Ave, Ann Arbor, MI 48109
zhng@umich.edu

MUSTER[1] is a single threading program/server developed for protein structure prediction. Compared to the widely-used sequence profile-to-profile alignment algorithm, MUSTER was designed to improve threading performance with multiple structural profile information. Except for the sequence profile match, five different structural features were added in MUSTER: (1) match of secondary structures of query and templates; (2) alignment of sequence-based query profile with structured-based template profile; (3) match of solvent accessibility of query and templates; (4) match of torsion angles (φ and ψ) between query and templates; (5) hydrophobic scoring matrix. In a benchmark test of 500 non-homologous proteins, it was found that the average TM-score[2] of the first threading alignment to native is nearly 5% higher than the PPA algorithm that was based on sequence profile-profile alignment and secondary structure fitting only[3]. The full-length models were constructed by MODELLER[4] based on the MUSTER template alignments. The MUSTER server, together with the program and structure library, is available for non-commercial users at http://zhanglab.ccmb.med.umich.edu/MUSTER.

1. Wu, S. & Zhang, Y. (2008) MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. *Proteins*.**72**, 547-556.
2. Zhang, Y. & Skolnick, J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins-Structure Function and Bioinformatics*.**57**, 702-710.
3. Wu, S. & Zhang, Y. (2007) LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic acids research*.**35**, 3375-3382.
4. Sali, A. & Blundell, T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*.**234**, 779-815.

OnD-CRF: Disorder prediction in proteins using Conditional Random Fields by optimizing the p-value cut-off

L. Wang[1,2] and U.H. Sauer[1,2]
[1] – *Department of Chemistry,* [2] – *Computational Life Science Cluster, CLiC, University of Umeå, Sweden*
uwe.sauer@chem.umu.se

An increasing number of proteins transfers key biological functions through intrinsically unstructured sequence intervals (Dunker, et al., 2002; Romero, et al., 1999). Finding the disordered regions in proteins will help to reduce bias in sequence similarity analysis, to identify protein domains boundaries and to guide structural and functional studies (Ferron, et al., 2006).
Order and Disorder prediction using Conditional Random Fields, OnD-CRF, is a machine learning method for accurate prediction of disordered amino acid intervals in proteins. The CRFs rely on features which are generated from the amino acids sequence and from secondary structure prediction and are able to take into account inter-relation information between two labels of neighboring residues. Benchmarking results based on CASP7 data rank the OnD-CRF model high within the fully automatic server group.

Methods

The training dataset used here contains 215,612 residues. But only 13,909 are defined as disordered (Cheng, et al., 2005a). The data set is derived from high-resolution crystal structures that lack coordinates for those amino acids that are considered to be disordered.
Performance is optimized with respect to the Area Under the ROC Curve, AUC, which is a measure of the overall predictor quality, with a value of 1.0 for a perfect predictor and 0.5 for a random predictor.
The OnD-CRF method makes use of the free program package CRF++ (http://crfpp.sourceforge.net/). The template file used for training the OnD-CRF model, contains the rules for generating the features which are extracted only from the protein sequence and the predicted secondary structure with the help of SSpro (Cheng et al., 2005b). We use cross-validation to find the optimal parameters for CRF++ and for the maximal AUC value.

Results

We use 10-fold cross validation and find that a sliding window size of nine amino acids optimizes the template file. The set of parameters which give rise to the best AUC value of 0.864 are: 1.018 for the hyper-parameter "C", which trades the balance between over-fitting and under-fitting and 5 for the parameter "f", which sets the cut-off threshold for the features. For all other parameters we use the default CRF++ 0.49 values.

As a result of the 10-fold cross validation, we find an optimal P-value cut-off of $P < 0.05$ for ordered and $P \geq 0.05$ for disordered amino acids(*). Using this cut-off the OnD-CRF model achieves an ACC of 0.790 based on the training dataset.

For benchmarking, we use the 96 targets from CASP7 and compare the OnD-CRF results to those of the fifteen methods that predicted 93 or more targets. Within the automatic server group, the OnD-CRF method reaches a very high overall performance, comparable to the best human expert methods such as ISTZORAN and fais. The results demonstrate, that our OnD-CRF method accurately predicts disorder in proteins in a fully automated way.

(*) Initially (from T0515 to T0536) we used a P-value cut-off of 0.05. CASP9 enforced the rules stating that the disordered state has probability scores >0.5. From T0537 on, we therefore multiplied the predicted disorder probability by 10 in order to move the discriminative line to 0.5. In those cases where the multiplied disorder probability values exceeded 1.0, we set them equal to 1.0.

Availability

OnD-CRF server: http://babel.ucmp.umu.se/ond-crf/

1. Dunker,A.K., Brown,C.J., Lawson,J.D., Iakoucheva,L.M. and Obradovic,Z. (2002) Intrinsic disorder and protein function, Biochemistry, 41, 6573-6582.
2. Romero,P., Obradovic,Z. and Dunker,A.K. (1999) Folding minimal sequences: the lower bound for sequence complexity of globular proteins, FEBS Lett, 462, 363-367.
3. Ferron,F., Longhi,S., Canard,B. and Karlin,D. (2006) A practical overview of protein disorder prediction methods, Proteins, 65, 1-14.
4. Cheng,J., Sweredoski,M.J. and Baldi,P. (2005a) Accurate Prediction of Protein Disordered Regions by Mining Protein Structure Data, Data Mining and Knowledge Discovery, 11, 213-222.
5. Cheng,J., Randall,A.Z., Sweredoski,M.J. and Baldi,P. (2005b) SCRATCH: a protein structure and structural feature prediction server, Nucleic Acids Res, 33, W72-76.

OnD-CRF-pruned

OnD-CRF-pruned: Disorder prediction in proteins using Conditional Random Fields by pruning the training data set

L. Wang[1,2] and U.H. Sauer[1,2]
[1] – *Department of Chemistry,* [2] – *Computational Life Science Cluster, CLiC, University of Umeå, Sweden*
uwe.sauer@chem.umu.se

The great challenge for protein disorder (Dunker, et al., 2002) predictors is to achieve accurate predictions even though the methods were trained on a very imbalanced data set. Commonly used training data sets contain only a few percent disordered residues and a large excess of ordered amino acids.

In order to alleviate the imbalance problem between ordered and disordered amino acids, we have implemented a novel strategy for training our **O**rder a**n**d **D**isorder (OnD) predictor that uses **C**onditional **R**andom **F**ields (CRFs) and relies on features that are generated from the amino acids sequence and from the predicted secondary structure.

In this method, called OnD-CRF-pruned, we prune the ordered regions in the sequences of the training data set in order to obtain a balanced training data set containing equal amounts of ordered to disordered amino acids. This approach enhances the accuracy of detecting disordered amino acids in proteins.

Methods

We use CRF++ 0.49 (http://crfpp.sourceforge.net/), to generate the OnD-CRF-pruned. From 10-fold cross validation, we find the optimal window size of nine amino acids which yields the best template file for the feature subset selection.

We train the OnD-CRF-pruned model on a dataset derived from high-resolution crystal structures. The dataset contains 215,612 residues, of which 13,909 are classified as disordered (Cheng, et al., 2005a).

The features for training the OnD-CRF-pruned model are extracted from the amino acid sequence and, using SSpro (Cheng, et al., 2005b), from the predicted secondary structure. Training the OnD-CRF-pruned model proved difficult since the training data set contains less than 6.5% of disordered amino acids, which leads to a label imbalance problem. In order to generate a balanced training dataset, we prune part of the ordered sequence intervals and keep only a limited number of ordered amino acids that flank the disordered regions. The pruned training dataset is, with a ratio of ordered to disordered residues of 1:1.04, almost perfectly balanced.

Results

For benchmarking, we use the 96 target structures available during CASP7 and compare the results obtained with OnD-CRF-pruned to the fifteen methods that predicted more than 93 target structures in CASP7. The performance of each method is then evaluated with respect to sensitivity, specificity, CASP Sscore, CASP Sproduct and ACC score.

The benchmarking results for all 16 disorder prediction methods show that, according to the CASP7 evaluation criteria, the OnD-CRF-pruned method scores very high within the automatic server group, and are comparable to human expert methods such as “ISTZORAN” and “fais”.

We believe that OnD-CRF-pruned is an accurate and effective method for the fully automated prediction of protein disorder.

Besides the prediction of disordered sequence intervals, we suggest that the accuracy of the OnD-CRF prediction can be used to determine domain boundaries for 3D structure analysis.

Availability

The OnD-CRF-pruned server: http://babel.ucmp.umu.se/ond-crf-p/

1. Dunker,A.K., Brown,C.J., Lawson,J.D., Iakoucheva,L.M. and Obradovic,Z. (2002) Intrinsic disorder and protein function, Biochemistry, 41, 6573-6582.
2. Cheng,J., Sweredoski,M.J. and Baldi,P. (2005a) Accurate Prediction of Protein Disordered Regions by Mining Protein Structure Data, Data Mining and Knowledge Discovery, 11, 213-222.
3. Cheng,J., Randall,A.Z., Sweredoski,M.J. and Baldi,P. (2005b) SCRATCH: a protein structure and structural feature prediction server, Nucleic Acids Res, 33, W72-76.

OPUS-Ma-sever

OPUS-MA-server

Yushao Cheng[(1)], Jianpeng Ma[(1,2)]
1- Department of Bioengineering, Rice University, Houston, TX
2- Verna and Marrs McLean Department of Biochemistry and Molecular Biology, Baylor college of Medicine, Houston, TX

Methods

OPUS-Ma-server uses a template-based method to predict protein tertiary structure. A template library is built from Protein Data Bank (PDB) by PISCES. The query sequence is aligned with each template in the library by a hybrid profile-profile alignment. Profiles include evolution –based profile (generated by PSI-BLAST) and other structure-based profiles. Alignment is done by implementing Smith-Waterman algorithm. MODELLER is used to generate final structure from the top scored template.

1. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res, 25(17):3389-402.
2. Eswar N, Marti-Renom MA, Webb B, Madhusudhan MS, Eramian D, Shen M, Pieper U, Sali A. (2006) Comparative Protein Structure Modeling With MODELLER. Current Protocols in Bioinformatics, John Wiley & Sons, Inc., Supplement 15, 5.6.1-5.6.30.
3. Jones DT. (1999) Protein secondary structure prediction based on position-specific scoring matrices. Journal of Molecular Biology. 292(2):195-202
4. Lindahl E and Elofsson A. (2000) Identification of related proteins on family, superfamily and fold level. Journal of Molecular Biology, 295(3): 613-625
5. Wang G and Dunbrack RL Jr. (2003) PISCES: a protein sequence culling server. Bioinformatics 19 (12): 1589-1591.
6. Zhang Y and Skolnick J. (2004) Scoring function for automated assessment of protein structure template quality. Proteins: Structure, Function, and Bioinformatics, 57 (4): 702–710
7. Zhang Y and Skolnick J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Research, 33(7):2302-2309

Panther_server

Fat-tailed error distributions, Delaunay triangulation and improving model quality

Robert W. Harrison
Georgia State University
rwh@gsu.edu

In CASP-9, the panther server generally was able to improve the quality of models, assessed by gdt and Q-score over its starting unrefined models. These improvements were typically about 0.03-0.05 in gdt, although there were examples that were both worse and better than this range. The RMSD also improved in about one half of the models. This is a dramatic change from earlier versions of the server, where the refinement process generally reduced model quality. The appropriate estimation and use of distance restraint information is a major reason for this change. The prior distribution for distances derived from experimental structure is not well characterized by a Gaussian, and the negative log(prior) penalty function used to enforce distance restraints was changed from the square error associated with Gaussian priors to a form associated with Cauchy distributions with highly useful results.

Methods

It is relatively straightforward to estimate distances between atoms to use as restraints in model building from a set of putatively homologous structures. The "unrefined" starting models for each homolog are generated for each reasonable alignment and then the distances between equivalent atoms in each model are identified and averaged. Similarly, it is possible to generate a library of structural fragments and by selecting the most similar fragment either from sequence or structure data, generate distance restraints for unobserved parts of the structure. However, naïve application of this approach results in poor quality models.

There are two likely reasons for this. First is the need to select distances that are truly determinative of the structure, as the selection of many degenerate distances results in a poorly conditioned problem where the optimal solution is hidden by the swarm of degenerate and contradictory data. Secondly, it is critical to estimate the reliability of these distance terms, and while standard deviation statistics are easy to use they tend to underestimate the error and are not reliable estimators of large outliers. It only takes a small number of highly erroneous distance restraints to destroy the quality of a model.

The panther server incorporates two novel techniques to incorporate this kind of distance information with resulting improvements in the accuracy of the refined model with respect to the starting model. The Delaunay triangulation of Ca-Ca, and Cb-Cb atoms is used to select the distances that are, at least in one geometrical sense, most critical in that they accurately describe the packing of the chain against itself in the folded structure. In addition to the distances along each Delaunay vector, the diagonal between them is also used. These distance terms can be summed over the best quality unrefined starting models, where model quality was predicted via a sequence alignment score based on Blossum weights (the alignment itself having been performed with a profile-profile algorithm), and the average values with small standard deviations are then used for restraints. Unfortunately, while large a priori standard deviations are good estimators of an unreliable distance restraint, a small a priori standard deviation is not a good estimator of reliability. Therefore a significant number of the estimated "good" distance restraints possessed large errors, which would dominate the model refinement and produce a poor model. Changing the assumed prior from Gaussian to Cauchy to handle a fat-tailed distribution (Cauchy

distributions do not possess a finite standard deviation, but are highly similar to Gaussian distributions for small errors) resulted in a system that enforced distance restraints that were reliable, while ignoring those that were not. Thus information about fold and model structure could be estimated over a wide range of homologous structures and transferred in a simple manner to the refinement of each model.

Alignments were generated using a profile-profile algorithm starting from profiles calculated with PSI-BLAST[1]. The Kullbeck entropy was used to estimate the likelihood that two profiles were identical and the scores were normalized in terms of standard deviations above the mean to control for composition effects. A minimum score cutoff of 2 standard deviations was used which allowed the occasional non-alignment to be further processed. If the initial pass of PSI-BLAST parameters failed to produce an alignment, the process was restarted with parameters adjusted to be more tolerant of distant homologs.

Models were built with the AMMP[2] program using a modified version of the current molecular mechanics potential where the dielectric was set to 80 and a supplemental hydrogen bond term was supplied. These changes help to preserve model quality, but were not in themselves capable of improving model quality in test systems. The distance restraints were changed from a Gaussian prior form $V = K(x - x_{target})^2 \approx -\log(e^{-k(x-x_{target})^2})$ where K is a force constant and x_{target} the target value to a Cauchy derived form $V = K\log(1 + m(x - x_{target})^2)$ where m is a constant adjusted to make the two priors equivalent for deviations of 1Å. The Cauchy form strongly restrains small deviations but ignores large ones.

Delaunay triangulation was used on each of the top 5 alignments, by calculating the unrefined models and then merging the distances from each. Restraints with high standard deviations were discarded. Structural clusters of 10mer's were used to supply structural data for insertions and missing fragments.

Results

The server did not always find a meaningful sequence alignment, but when it found one the quality of the model was improved from the starting point in almost every case (as of 9/17/2010 only two models had slightly lower quality after refinement). The changes were small, of the order of 0.03 in GDT, but consistent, indicating that the refinement process actually was working. The RMSD also improved in a significant number of models.

Availability

The server is available at http://bmcc3.cs.gsu.edu.

1. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
2. Harrison, R.W., Chatterjee D., and Weber I.T. "Analysis of six protein structures predicted by comparative modeling techniques." (1995) Proteins: Structure Function and Genetics 23:463-471.

Phyre2

Simulated protein synthesis and folding with template-derived distance constraints in Phyre2

L.A. Kelley[1], B.R. Jefferys[1] and M.J.E. Sternberg[1]
[1] – *Imperial College London*
l.a.kelley@imperial.ac.uk

Phyre2 (http://sbg.bio.ic.ac.uk/phyre2) is an automated method for the prediction of protein 3D structure combining *de novo* and template-based methods using a dynamic model of protein synthesis and folding. Template recognition and modeling is performed as in the previous method used at CASP8 (*Phyre de novo*)[1]. These individual models are used to derive distance constraints for use in a modified version of our *de novo* folding technique, Poing[2].

Methods

A protein sequence is initially scanned against a 50% non-redundant sequence database (Uniref50) using PSI-Blast[3] followed by secondary structure prediction using PSI-pred[4]. A hidden Markov model of the sequence is generated and scanned against a library of HMMs using the HHsearch 1.5.1 package[5]. High scoring templates are chosen to simultaneously maximize coverage of the input sequence and confidence in the homology. These templates are then used to build a small number (usually <10) of simple alpha carbon-only models.

Each of these models is used to generate a set of pairwise distances between residues in space. These distances are converted into simple springs within a modified version of the Poing[2] *de novo* modeling tool. Poing then slowly synthesizes the protein from a virtual ribosome, adding distance springs as more residues are added to the growing chain. Insertions and large missing regions are modeled using the Poing *de novo* protocol. The Poing simulation is repeated between 5 and 100 times depending on factors such as protein length, beta structure content and template coverage.

In proteins for which either no confident templates are found or for which template coverage is very low, short fragment models are constructed across the length of the input sequence using low confidence hits from HHsearch. These fragments are used to generate the distance springs and the poing simulation is run 100 times. Finally, the resulting models are clustered and the model with the greatest similarity to all other models in the pool is chosen.

The full protein backbone is then reconstructed using Pulchra[6] and sidechains are placed using our in-house version of the R3 sidechain placement algorithm[7].

Availability

Phyre2 is available at: http://sbg.bio.ic.ac.uk/phyre2.

1. Kelley,L.A. and Sternberg,M.J.E. (2009). Protein structure prediction on the web: a case study using the Phyre server. *Nature Protocols.* **4**, 363-371.
2. Jefferys,B.R., Kelley,L.A. and Sternberg, M.J.E. (2010). Protein folding requires crowd control in a simulated cell. *J. Mol. Biol.* ***397***, 1329-1338.
3. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
4. Jones,D.T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* ***292***, 195-202.
5. Söding J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**, 951-960.

6. Rotkiewicz P., Skolnick, J. (2008). Fast method for reconstruction of full-atom protein models from reduced representations *J. Comp. Chem.* **29**, 1460-1465.
7. Xie,W. and Sahinidis,N.V. (2006) Residue-rotamer-reduction algorithm for the protein side-chain conformation problem. *Bioinformatics* **22**, 188-194.

PLATO

Fully Automated Structure Prediction using Ideal Forms

M.I. Sadowski[1], K. Maksimiak[1], J.T. Macdonald[1,2,3] and W.R. Taylor[1]

[1] - Division of Mathematical Biology, MRC National Institute for Medical Research, The Ridgeway, Mill Hill, London, NW7 1AA UK.

[2] - Centre for Synthetic Biology and Innovation, Imperial College London, London, SW7 2AZ, UK

[3] - Division of Molecular Biosciences, Imperial College London, London, SW7 2AZ, UK

msadows@nimr.mrc.ac.uk

A server for *ab initio* protein structure prediction using ideal forms[1] was developed. This was essentially a fully-automated version of the previously described build method [2] however it included refinements to enable a fully-automatic solution and improve model selection and ranking.

Methods

The first step was to identify potential structural domains by aligning query sequences to the GENE3D hidden Markov model library [1] using HMMer3 [2] with default parameters. Counts of alignment endpoints were taken for each sequence position and smoothed over five iterations by taking the average over a seven-residue window, ignoring the first and last ten residues. Potential splits were recorded as the locations of maxima after smoothing. Final predictions were generated using a greedy algorithm which attempts to split the sequence into regions no shorter than 75 residues and no longer than 300.

Profiles for target sequences were generated by alignment to a weekly-updated local copy of the NR database using PSIBLAST [1] following which the alignment was culled to a small number of representatives. Predictions of secondary structure were made using two methods (PSIPRED [2] and YASPIN [3]) for all representatives in the alignment. Predictions were grouped and converted to element-level predictions by testing all possible alternatives for ambiguous elements: present/absent for short elements (length < 3 for strands, length < 4 for helices) and helix/strand for ambiguous regions.

For each prediction all compatible ideal forms were identified and used as templates for prediction. A given form provides a lattice representation for an arrangement of secondary structures in either a three-layer alpha/beta/alpha, four-layer alpha/beta/beta/alpha or polyhedral all-alpha arrangement. Possible topologies were generated for each lattice by generating all permutations compatible with lengths of predicted loops and sequence hydrophobicity. In a novel step the choice of lattice was filtered by comparison with known SSE sequences using BLAST. A population of hundreds of alpha-carbon models was generated in this way for each domain.

C-alpha models were compared to predictions for secondary structure content and models with incorrect numbers of elements or which were not compact were removed. Alpha-carbon positions were refined and backbones added using PRODART [8] following which a second filter was applied based on Rosetta[9] and DFire[10] energy scores, filtering the high-energy 50% of both sets. Finally models were ranked by DFire scores and the top five were returned. Sets of domain predictions were assembled using MODELLER [11]

1. Lees,J., Yeats,C., Redfern,O., Clegg,A., Orengo,C. (2010) Gene3D: merging structure and function for a thousand genomes. *Nucleic Acids Research.* **38**, D296-D300
2. http://hmmer.wustl.edu
3. Taylor,W.R. (2002) A ‘periodic table’ for protein structures. *Nature***416,** -660

4. Taylor,W.R., Bartlett,G.J., Chelliah,V., Klose,D., Lin,K., Sheldon,T. & Jonassen,I. (2008) Prediction of protein structure from ideal forms. *Proteins***70**, 1610-1619
5. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.***25**, 3389-3402.
6. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* ***292***, 195-202.
7. Lin,K., Simossis,V.A., Taylor,W.R. and Heringa,J. (2005) A Simple and Fast Secondary Structure Prediction Algorithm using Hidden Neural Networks. *Bioinformatics*. **21**, 152-9.
8. Macdonald,J.T., Maksimiak,K., Sadowski,M.I., & Taylor,W.R. (2010) *De novo*backbone scaffolds for protein design *Proteins***78**, 1311-1325
9. *Das R,Baker D. (2008) Macromolecular modeling with rosetta. Annu Rev Biochem* ***77****, 363–382.*
10. Yang T., Zhou Y. (2008) Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions *Protein Science***72**, 1212-1219
11. Sali,A. & Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol*. **234**, 779-815

POEM

Performance of an All-Atom Free-Energy Approach for Protein Structure Prediction

[1]P. Anand, [1]T. Strunk, [1]I. Meliciani, [2]M. Brieg, [1]M. Wolf, [2]K. Klenin and [1]W. Wenzel
[1]Institute for Nanotechnology, [2]Steinbuch Centre for Computing,
Karlsruhe Institute of Technology, Germany
wolfgang.wenzel@kit.edu
http://www.research.kit.edu/biostruct

De novo prediction of protein tertiary structure on the basis of amino acid sequence remains one of the outstanding problems in biophysical chemistry. We have developed an all-atom free energy forcefield PFF02 which stabilizes a wide array of proteins. Recently we have implemented these techniques in POEM@HOME (http://boinc.fzk.de), a world-wide distributed computational architecture.

Methods

In CASP9, we participated as a human expert group, where the type and extent of human input varied depending on the target difficulty. For a few cases for which closely related sequence homologs were available and sequence alignment was possible, we performed homology modeling. For other few targets, the best template (from PSI-Blast, 3D-Jury[1], Phyre[2]) was selected using a combination of automatic server (domain prediction and secondary structure alignment) and manual inputs. The best template was then used for construction of the optimal sequence alignment using ClustalW/Tcoffee, while additional tools (MOE/Modeller) were used for converting these sequence alignments into 3D models. The resulting models were subsequently relaxed in refinement simulations using POEM@HOME, as described below. For complicated targets for which automated servers (Blast, Phyre) could not reliably detect any sequence homologs (templates) with significant e-values, we used a fragment based approach.

Refinement Protocol: All structures were refined using simulated annealing simulations in the POEM++ protein modelling framework and the PFF02 forcefield, a semi-empirical all-atom forcefield with combined hydrogens, proven to stabilize a multitude of different folds[3]. Automated structure refinement was carried out using fixed bond-lengths and variable main-chain and side-chain dihedral angles. In addition to energetic relaxation, refinement of the templates comprised of the addition of missing residues and guidance towards secondary structure constraints as predicted by the PSIPRED Server. Disulphide bridges were modelled by imposing distance constraints.

Fragment Model Protocol: Initial decoy sets for the fragment based modeling scenario were generated using the Rosetta 3.1 software suite[4] and the default ab-initio protocol. Depending on the available resources a set of 10.000 - 15.000 structures was generated, clustered to avoid redundant relaxations and relaxed on the POEM@HOME infrastructure. PFF02 best-energy structures were then submitted after another possible refinement step.

Results

One exemplary protein structure, where this modeling protocol was applied was T0537. Possible template structures for this model were 1K0E and 1DBG. An alignment between the target and 1K0E resulted in an overall realistic global dimer-like fold, with a beta sheet core isolated circular by helices. 1DBG on the other hand resulted in a completely different global fold, a beta-sheet-only tube. Human inspection in our group favored the 1K0E model. The gene-family of T0537 and 1DBG matched however, leaving us undecided, which model to submit. Energy relaxation selected the 1DBG model as

the best-energy model by a wide margin (~40 kcal/mol difference), which corresponded to the correct global fold. The relative scores calculated with the TMScore program are GDT-TS score: 0,68 and RMSD: 3,5 Angstrom.

Availability

POEM++ is still in development. Please contact our group for an evaluation copy.

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. Kelley L.A., Sternberg MJE (2009) “Protein structure prediction on the web: a case study using the Phyre server” Nature Protocols 4; 363 - 371
3. Verma A., Wenzel W., (2009) “A free-energy approach for all-atom protein simulation” Biophys. J. 96;83-3494
4. Kim D.E., Chivian D., Baker D. (2004) “Protein structure prediction and analysis using the Robetta server.” Nucleic acids research Jul, 32 W526-31

prdos2

De novo protein tertiary structure prediction server accelerated by GPU computing techniques

Takashi Ishida[1,2], Yutaka Akiyama[1]
[1] -Graduate School of Information Science and Engineering, Tokyo Institute of Technology , [2] - CompView, Global COE
t.ishida@bi.cs.titech.ac.jp

The prdos2-server is an automated protein tertiary structure and disorder prediction server. For tertiary structure prediction, the server tried to identify the structural templates of a target sequence by using a fold recognition technique. Then the server applied *de novo* prediction based on the fragment assembly method if there was no significant template. Generally, *de novo* prediction requires large computational resources. Thus, we implemented our *de novo* prediction system on the Graphics Processing Units (GPUs) and accelerated the conformational space samplings.

Methods

The server tried to identify the structural templates of a target sequence for the PDB by HMM-HMM comparison using HHsearch program[1]. If statistically significant templates were found, tertiary structure models were generated by using Modeller program based on those templates. If there were some long unaligned regions in the alignment, those regions were modeled by our *de novo* structure prediction system described later. Finally, the models were ranked according to our statistical potentials.

If reliable alignments could not be found for the whole target sequence, the server generated tertiary structure models by using our *de novo* modeling system based on the fragment assembly method. The system searched conformational spaces by simulated annealing method using a potential energy function including terms of potential based on contact number prediction[2], atom clashes, and hydrogen bonding. However, this sampling process requires a lot of computation. Thus, we implemented this process on the GPUs by using NVIDA CUDA toolkit version 2.3. Although GPUs are designed specifically for computer graphics and thus are very limited in terms of operations and programming, they can operate massive parallel jobs much faster than CPUs because they have hundreds of streaming multi-processors in the core. Our GPU-accelerated system on the Tesla C1060 GPU achieved a speedup of up to 3.4 times with respect to a single CPU core. By using high-computational power by GPUs, the server produced more than 100,000 models for each target, and selected five prediction models by using the potential energy and structural clustering techniques[3]. Finally, side chain modeling was performed by using SCWRL version 3.0[4].

1. Soding J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**(7), 951-960.
2. Ishida T, Nakamura S, Shimizu K. (2006) Potential for assessing quality of protein structure based on contact number prediction. *Proteins* **64**(4): 940-947.
3. T. Ishida, T. Nishimura, M. Nozaki, T. Inoue, T. Terada, S. Nakamura, K. Shimizu (2003) Development of an ab initio protein structure prediction system ABLE, *Genome Informatics*, **14**, 228-237
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.

PRECORS
PRECORS-QA
FEIG

Protein structure prediction with quality assessed scoring and simulation-based refinement

M. Feig and K. Vadivel
Department of Biochemistry & Molecular Biology, Michigan State University
feig@msu.edu

Protein structure prediction is accomplished with a multi-step protocol where initial models are generated with a variety of different methods and subsequent scored. During the scoring stage multiple scoring functions are applied and a new statistics-based criterion is applied to determine the most reliable scoring function for a given set of decoys (PRECORS and PRECORS-QA servers). In the human category (FEIG) best-scoring models from server predictions are ranked and refined with simulation-based methods. Finally, in the refinement category, extended atomistic simulations are carried out along with target loop resampling to refine the given initial models.

Methods

For server predictions we generated initial with a number of different methods, including comparative models from MODELLER based on single and multiple template alignments from PDB-BLAST, HHPRED[1], FFAS[2], SAM-T02, and PROSPECTOR[3] as well as models generated by TASSER[3] and an in-house implementation of I-TASSER[4]. All of the generated models were then scored with DFIRE[5], OPUS-PSP[6], DOPE[7], and a series of pairwise and multibody potentials compiled by the Jernigan group[8]. The reliability of each scoring function was then assessed with a recently developed scoring confidence index[9] and the scoring method(s) predicted to be performing best was used to score the models.

Human predictions were generated by selecting the best server predictions as starting models and carrying out refinement with replica-exchange simulations using either atomistic force fields or the coarse-grained PRIMO model[10]. Sampling in the refinement simulations were restrained to remain within a few Å from one or more initial models.

Predictions in the refinement category were generated either with extensive loop resampling followed by the scoring protocol outlined above or by long molecular dynamics simulations in explicit solvent with a modified version of the CHARMM force field to improve the sampling of backbone torsion angles.

Results

Results will be available after fully assessing our submissions.

Availability

Most of the components used in our structure prediction pipeline are available publicly. The actual pipeline will be made available in form of scripts if it proves to be successful.

1. Soding, J., A. Biegert, and A.N. Lupas, *The HHpred interactive server for protein homology detection and structure prediction.* Nucleic Acids Research, 2005. **33**: p. W244-W248.

2. Jaroszewski, L., et al., *Fold prediction by a hierarchy of sequence, threading, and modeling methods.* Protein Science, 1998. **7**(6): p. 1431-1440.
3. Zhang, Y., A.K. Arakaki, and J.R. Skolnick, *TASSER: An automated method for the prediction of protein tertiary structures in CASP6.* Proteins-Structure Function and Bioinformatics, 2005. **61**: p. 91-98.
4. Wu, S., J. Skolnick, and Y. Zhang, *Ab initio modeling of small proteins by iterative TASSER simulations.* BMC Bioinformatics, 2007. **5**: p. 17.
5. Zhou, H.Y. and Y.Q. Zhou, *Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction.* Protein Science, 2002. **11**(11): p. 2714-2726.
6. Lu, M.Y., A.D. Dousis, and J.P. Ma, *OPUS-PSP: An orientation-dependent statistical all-atom potential derived from side-chain packing.* Journal of Molecular Biology, 2008. **376**(1): p. 288-301.
7. Shen, M.Y. and A. Sali, *Statistical potential for assessment and prediction of protein structures.* Protein Science, 2006. **15**(11): p. 2507-2524.
8. Feng, Y.P., A. Kloczkowski, and R.L. Jernigan, *Potentials 'R'Us web-server for protein energy estimations with coarse-grained knowledge-based potentials.* BMC Bioinformatics, 2010. **11**: p. -.
9. Zavodszky, M.I., et al., *Scoring Confidence Index: Statistical Evaluation of Ligand Binding Mode Predictions.* Journal of Computer-Aided Molecular Design, 2009. **23**: p. 289-299.
10. Gopal, S.M., et al., *PRIMO/PRIMONA: A coarse-grained model for proteins and nucleic acids that preserves near-atomistic accuracy.* Proteins: Structure, Function, and Bioinformatics, 2010. **in press**.

PRMLS

CASP9 protein structure modeling using computational methods and human expertise

Jimin Pei
Howard Hughes Medical Institute, The University of Texas Southwestern Medical Center at Dallas
jpei@chop.swmed.edu

Template identification: Templates were identified by PSI-BLAST [1] and HHsearch [2]. Top templates were manually inspected and selected.
Target-template alignments: The alignments were generated by PROMALS3D [3] followed by manual curation.
Model construction: Initial models for the targets were generated by MODELLER [4] with input target-template(s) alignments. Loop regions were manually defined and modeled by MODELLER or ROSETTA [5]. For some targets, structural refinements were done by ROSETTA. For targets without suitable templates, ROSETTA *ab initio* predictions were made.
Results: Human expertise was successful in selecting distantly related templates for some difficult targets (e.g., T0537, T0621, T0568, and T0571 first domain). Even in these cases, the structural variations between targets and templates prevent good modeling results (the targets are intrinsically difficult). When a wrong template was forced (e.g., T0531, T0574, T0606, T0624 and T0571 second domain), predictions were completely wrong in terms of overall structural fold. The lesson is that *ab initio* predictions should be used in these cases when clear homology relationships between a target and the potential template(s) cannot be established. In one case (T0581), *ab initio* prediction correctly identified the overall fold.

1. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
2. Soding, J., *Protein homology detection by HMM-HMM comparison.* Bioinformatics, 2005. **21**(7): p. 951-60.
3. Pei, J., B.H. Kim, and N.V. Grishin, *PROMALS3D: a tool for multiple protein sequence and structure alignments.* Nucleic Acids Res, 2008. **36**(7): p. 2295-300.
4. Sali, A., et al., *Three-dimensional models of four mouse mast cell chymases. Identification of proteoglycan binding regions and protease-specific antigenic epitopes.* J Biol Chem, 1993. **268**(12): p. 9023-34.
5. Rohl, C.A., et al., *Protein structure prediction using Rosetta.* Methods Enzymol, 2004. **383**: p. 66-93.

ProC_S3

ProC: Residue-Residue Contact Predictions Using Random Forest Models

J.W. Fang, Y.Q. Li and S. Gao
[1] - *Applied Bioinformatics Laboratory, the University of Kansas, 2034 Becker Dr., Lawrence, KS 66049.*
jwfang@ku.edu

The ability to accurately predict residue-residue contacts can be useful in protein structure predictions, especially for template-free modeling[1]. Here we present a Random Forest model for predicting residue-residue contacts. This model was constructed using 1,287 sequence-based features. The model was trained and cross-validated using a set of 1,490 non-redundant proteins with high resolution structures. The model was then tested in a set of 329 non-redundant proteins, all with sequence similarity less than 25% to the chains in the training dataset. The predictive model, implemented in the server ProC_S3, was an updated version of two CASP8 servers (Group names: RR_Fang_1 and RR_Fang_2). For comparison, these two servers also participated in CASP9 (renamed as ProC_S1 and ProC_S2 for consistence).

Methods

The predictive models were built using the Random Forest algorithm[2]. The Random Forest algorithm is an ensemble technique that utilizes many independent decision trees to perform classification or regression. Each of the member trees is built on a bootstrap sample from the training data including a random subset of available variables. The Random Forest algorithm is robust and particularly suitable for classifying high-dimensional and noisy data.

For the current model, we assembled a set of 1287 features for long-range contact predictions and 1282 features for short and medium contact predictions. These features can be roughly grouped into four categories: local window features, pairwise information features, in-between segment features, and residue properties of whole proteins and in-between segments. In addition to commonly used features such as position specific scoring matrices, we introduced a number of new features in the model. For example, we developed a novel protein alphabet with seven types of residues based on a large-scale statistical analysis and clustering study.

Results

In the blind benchmark test, the model delivered an accuracy of 30.4% and coverage of 4.2% for the top L/5 long-range (e.g. sequence separation >= 24) predictions. For the 121 targets in CASP8, the average accuracy reached 33.3% with an average coverage of 5.6%. Preliminary analysis based on 76 CASP9 targets showed that the average accuracy of the top L/5 long-range contact predictions was improved from 25.4% and 21.7% of ProC_S1 and ProC_S2, respectively, to 28.9% of ProC_S3. The average coverage was increased from 4.03%, 3.38% to 4.51%.

Availability

All current and previous ProC servers are freely available at http://www.abl.ku.edu/Pred_CMAP/.

1. Sitao,W., & Yang,Z. (2008) A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. *Bioinformatics*, ***24***, 924-931
2. Breiman L. (2001) Random Forests. *Machine Learning*, ***45***, 5-32.

ProQ2

Model Quality Assessment and Ranking using ProQ2

Arjun Ray[1] and Björn Wallner[1]
[1] - *Dept.Biochemistry and Biophysics, Stockholm University, Center for Biomembrane Research, Swedish e-Science Research Center*
bjorn@cbr.su.se

ProQ2 is non-consensus based Model Quality Assessment Program, i.e. it does not use any other information than found in the model. ProQ2 calculates structural properties from the model and use a SVM to predict the quality. The structural properties are similar to the ones used in the previous version of ProQ[1], but additional features that were shown to improve performance have been added. ProQ2 was used as an MQAP for model selection not only by us but also for many of the methods used by the Elofsson group. We participated with ProQ2 and the previous version ProQ (as reference) in the TS, QA and TR categories. For the TS category we used it to rank models from the Pcons.net server[2] and in the TR category we ranked and selected models refined using the standard relax protocol in Rosetta[3]

Methods

ProQ2 uses a combination of the following features: atom–atom and residue–residue contacts, surface accessibility, secondary structure and evolutionary information to prediction both local and global residue quality. All features are calculated over a sequence window and the local structural quality as measured by S-score is predicted using and SVM. S-score is calculated using the following formula S-score ($S_i=1/(1+(d_i/3)^2)$), where d_i is the distance to the correct structure in the superposition that maximizes the sum of S_i. A global score to be used for ranking models is obtained by summing up the local scores and divide by the target length.

Many of the features used in ProQ2 are similar to the ones used in ProQ, but there are some key differences:

- All structural statistics, like atom-atom contacts, residue-residue contacts and surfaces, calculated from the models are weighted using sequence profiles to smoothen the training data and also make the overall method less sensitive to minor sequence changes.
- Predicted secondary structure is encoded differently.
- Predicted surfaces are added.
- Sequence profiles are used directly.
- Position based conservation is also used to give less weight to less conserved positions.
- Global parameters, such as overall agreement between predicted and actual secondary structure, and agreement between predicted and actual exposed surfaces, were used also for the local prediction.

Results

The qualities of the highest ranked models for each target seem to be the highest among non-consensus methods with a performance similar to a standard consensus based protocols like Pcons[4]. The

improved performance is a consequence of tiny contributions from each new feature to the total performance with the largest performance increase obtained by including global parameters in the local prediction features. Preliminary results on the 95 CASP9 targets currently available, shows a significantly higher GDT_TS of the first ranked models by ProQ2 compared to ProQ (+~15%) and significantly better correlation.

Availability

ProQ2 will be available as standalone program and web server at http://proq2.cbr.su.se.

1. Wallner, B. and Elofsson, A. (2003) Can correct protein models be identified? Protein Sci 12 (5) : 1073-1086.
2. Wallner, B., Larsson, P. and Elofsson, A. (2007) Pcons.net: protein structure prediction meta server. Nucleic Acids Res 35 (suppl_2) : W369-W374
3. Rosetta 3.1, http://www.rosettacommons.org/.
4. Wallner, B. and Elofsson, A. (2006) Identification of correct regions in protein models using structural, alignment, and consensus information. Protein Sci 15 (4) : 900-913.

pro-sp3-TASSER

Pro-sp3-TASSER server for protein structure prediction in CASP9

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

We have updated the pro-sp3-TASSER server[1] for protein structure in CASP9 with an SVM-based template ranking and a new model ranking method FTCOM[2] for medium/hard targets. We also implemented a new procedure for building all atom models from the C□-only models from TASSER[3] simulations.

Methods

Pro-sp3-TASSER server uses a threading method with five independent component scores: one component is taken directly from SP[3] [4] threading and the other four components are modified from the PROSPECTOR_3 [5] method. Targets are classified by their SP[3] threading Z-score into Easy, Medium and Hard categories. For Medium/Hard targets, the top 200 templates from each component score are re-ranked by an support vector machine (SVM) method. Furthermore, alternative alignments are generated by a parametric approach and good alignments are then selected by TASSER-QA[6]. The top templates identified by each threading score (after applying the SVM) along with their alternative alignments are combined to derive contact and distant restraints for model refinement by short TASSER simulations. For Medium/Hard targets, chunk-TASSER[7] is also used to generate full-length models. Multiple short TASSER or chunk-TASSER runs are used to generate an ensemble that has up to 150 full-length models. Subsequently, the top 20 models are selected from this ensemble by FTCOM. These are used to generate contact and distance restraints for longer TASSER simulations. Special attention is paid to possible multiple domain targets. We check the coverage of the top template as identified by its SP[3] score; if more than 50 residues are unaligned, the unaligned and aligned regions are modeled separately in addition to modeling the full length target sequence. The separately modeled, possible domains are then overlapped onto the full-length models in the second round of TASSER refinement. Other special cases are when the Z-score of the first SP[3] template is 2.0 units higher than the second template or when a single template has > 50% sequence identity to the target; then, only models from the first or the single high sequence identity template are used in TASSER simulations. Final models are selected from both rounds of TASSER runs by FTCOM. Ideal geometry backbone models are built from those selected Cα-only cluster centroid models. An in-house template-based side-chain building procedure was employed to build the side-chains of submitted models.

Results

Pro-sp3-TASSER server models have better geometry and better H-bond score and side-chain accuracy compared to CASP8 predictions according to our benchmark test. It is still among the top predictors, especially for human/hard targets, according to unofficial assessment at http://zhanglab.ccmb.med.umich.edu/casp9/.

Availability

Pro-sp3-TASSER program and service are available through our webpage at

http://cssb.biology.gatech.edu/

1. Zhou,H and Skolnick, J. (2009) Protein structure prediction by pro-sp3-TASSER. Biophysical Journal. **96**, 2119-27.
2. Zhou,H and Skolnick, J. (2010) Improving threading algorithms for remote homology modeling by combining fragment and template comparisons. Proteins. **78**, 2041-8.
3. Zhang, Y. and J. Skolnick (2004) Automated structure prediction of weakly homologous proteins on genomic scale. Proc. Natl. Acad. Sci. (USA) **101**,7594--7599.
4. Zhou, H. and Zhou,H. (2005) Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins **58**, 321--328.
5. Skolnick, J., D. Kihara, and Y. Zhang (2004) Development and large scale benchmark testing of the PROSPECTOR 3.0 threading algorithm. Proteins **56**, 502--518.
6. Zhou,H. and Skolnick,J.(2007) Protein model quality assessment prediction by combining fragment comparisons and a consensus Cα contact potential. Proteins **71**,1211--1218.
7. Zhou, H and Skolnick, J. (2007) Ab initio protein structure prediction using chunk-TASSER. Biophysical Journal. **9**3,1510-8.

PROTAGORAS

Fully Automated Structure Prediction using Template-based modelling and Ideal Forms

M.I. Sadowski[1] and W.R. Taylor[1]

[1] - *Division of Mathematical Biology, MRC National Institute for Medical Research, The Ridgeway, Mill Hill, London, NW7 1AA UK.*

msadows@nimr.mrc.ac.uk

A hybrid server for automated protein structure modelling was developed using profile-sequence and profile-profile alignment methods for the template based modelling and the build method for structure prediction using ideal forms for *ab initio* prediction of regions not covered by templates.

Methods

Profiles for target sequences were generated by alignment to a weekly-updated local copy of the NR database using PSIBLAST [1]. Profiles were then used to search the PDBAA sequence database generated by the PISCES server [2] and generate predictions of secondary structure using PSIPRED[3]. Multiple alignments generated by PSIBLAST were filtered to remove sequences covering <75% of the length of the query and columns containing > 99% gaps, following which they were used to generate hidden Markov models and search PDB70 with HHPred (v1.50)[4].

Templates were chosen by E-value (lowest first) to maximise coverage and minimise overlap. Remaining uncovered sections were modelled using three- and four-layer alpha/beta/alpha forms and polygonal all-alpha forms by assigning SSE lattice frameworks compatible with PSIPRED and YASPIN [5] secondary structure and generating compatible topologies [6]. Models were ranked by hydrophobicity and compactness. In each case only a single model was chosen. Final assemblies of template models and *ab initio* predictions were generated using MODELLER 9v3 [7].

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. Wang,G. & Dunbrack Jr,R.L.. (2003) PISCES: a protein sequence culling server. *Bioinformatics* **19**, 1589-1591
3. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol. **292***, 195-202.
4. Soding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**, 951-960
5. Lin,K., Simossis,V.A., Taylor,W.R. and Heringa,J. (2005) A Simple and Fast Secondary Structure Prediction Algorithm using Hidden Neural Networks. *Bioinformatics*. **21**, 152-9.
6. Taylor,W.R., Bartlett,G.J., Chelliah,V., Klose,D., Lin,K., Sheldon,T. & Jonassen,I. (2008) Prediction of protein structure from ideal forms. *Proteins* **70**, 1610-1619
7. Sali,A. & Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol*. **234**, 779-815

ProteinShop

Structure Prediction of Beta Proteins Using BuildBeta

Scott Refugio[1], Daniel Cheung[1], Chengcheng Hu[1], Nelson Max[1,2] and Silvia Crivelli[1,2]
[1]Dept. of Computer Science, Univ. of California, Davis, CA 95616, [2]Lawrence Berkeley Laboratory, Berkeley, CA 94720
SNCrivelli@lbl.gov

We describe an ab-initio method to predict the tertiary structure of beta proteins. This method uses a two-phase approach: first, it thoroughly samples the protein conformation space using secondary structure predictions; then, it ranks models using a combination of different model-evaluation functions. The analysis of the results suggests that this method generates good models although imperfect predictions of secondary structure influence its accuracy. However, further research is needed to develop a more reliable scoring function.

Methods

Given the protein primary sequence of amino acids and secondary structure predictions, our method can automatically and thoroughly sample a protein conformation space. This is particularly important for predicting beta proteins, because of the complexity of sampling long-range strand pairing. Using some basic packing principles, inverse kinematics, and β-pairing scores, this method creates β-sheet arrangements using secondary structure predictions based on a number of secondary structure prediction servers. The method samples in the vicinity of the native fold, assuming correct or nearly correct secondary structure predictions are given.

Our method is based on ProteinShop1, a protein-modeling program that enables users to intuitively create models of a protein of interest using algorithms borrowed from computer games and animation, and also provides tools for automatic model creation. Among those tools is BuildBeta2, which creates all possible β-sheet arrangements given a prediction file containing the sequence of amino acids and secondary structure predictions.

Phase I: This phase consists of the following steps. First, we use the BioInfoBank meta-server3, which uses the 3D-jury consensus approach4 to select those targets that are likely to have new folds. Second, we create one or more consensus secondary structure prediction files according to the secondary structure predictions from various servers.5-9 BuildBeta reads those prediction files and assigns ideal values to the backbone dihedral angles of residues predicted to be α-helices and β-strands. Then, this extended conformation is folded into β-sheet arrangements using inverse kinematics algorithms that treat α-helices and β-strands as rigid bodies and adjusts the flexible backbone in the coil regions and that ultimately decide whether an arrangement is feasible or not. BuildBeta uses sequence-matching specificity10 to align the strands to form hydrogen bonds. Once the sheets are formed, BuildBeta places α-helices at suitable positions parallel to the constructed β-sheets to avoid the collision between secondary structure motifs as well as to bury hydrophobic residues.

BuildBeta only focuses on changing dihedral angles to achieve different conformations and does not use an energy function in the process. Once the structures are generated, BuildBeta optimizes the rotamer choices for the side chains by running the SCWRL411 program, which has been added to ProteinShop as a plug-in. In addition, all the structures generated are locally minimized using the Amber force field12 and the BFGS algorithm implemented in Gromacs13.

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

We use a combination of different scoring functions: Dfire14, Ramp15 and Crysol16. These three scores focus on different features and they usually do not agree with each other based on previous experiences. However, it is possible that there are some internal correlations between these three scores and we have combined them to obtain a better score for ranking the models. We use a neural network to build the implicit correlations among the three scores, and the resulting score, although far from perfect, seems to be better and more general. Because we only use this combination score for ranking structures created by BuildBeta, we use a decoy set composed of structures generated by our modeling method as the training set. For each protein, we generate a set of different models and remove the models that are not compact or have a collision score bigger than a threshold. Then, we score these models using Dfire, Ramp and Crysol as well as compute the RMSD to the native structure. To deal with the different ranges of scores and RMSD, we normalize the scores and RMSD to be in the interval [0,1]. Then, for each protein, we have a training data that has three inputs: the Dfire, Ramp and Crysol scores, and one output: the RMSD. We train our neural network with half of the training data and test it with the other half of data. Our testing results show that the output from the neural network has an average correlation value around 0.8 with the real RMSD (normalized into [0.1]). Finally, we select five models according to the combination score and human intuition with convenient interactive operations implemented in ProteinShop.

The Parallel Approach: The most time-consuming part of the computation is applying inverse kinematics to the coils to build all possible protein structures with proposed topologies and alignments. This process can take many hours or even days as shown in [2]. We have sped up this process substantially by working on different sets of topologies in parallel, which has allowed us to complete proteins with up to 10 predicted β-strands in 1 day running on 250 processors.

Acknowledgements

The authors wish to thank NERSC for valuable computing hours on their clusters. Many thanks also to the CS Dept. at UC Davis for the use of their cluster.

1. CrivelliS., KreylosO., HamannB., MaxN. & BethelW. (2004) ProteinShop: A tool for interactive protein manipulation and steering. Journal of Computer-aided Molecular Design. 18, 271-285.
2. MaxN., HuC., KreylosO., and CrivelliS. (2009). BuildBeta-A system for automatically constructing beta sheets. Proteins: Structure, Function, and Bioinformatics, 78(3), 559-574, DOI: 10.1002/prot.22582.
3. http://meta.bioinfo.pl/submit_wizard.pl.
4. GinalskiK., ElofssonA., FischerD., & RychlewskiL. (2003) 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics. 19(8),1015-1018.
5. McGuffinL.J., BrysonK. & JonesD.T. (2000) The PSIPRED protein structure prediction server. Bioinformatics 16, 404-405.
6. http://compbio.soe.ucsc.edu/SAM_T08/T08-query.html.

7. http://www.predictprotein.org/main.php.
8. http://www.compbio.dundee.ac.uk/www-jpred/.
9. http://www.meilerlab.org/web/.
10. ZhuH. & BraunW. (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.
11. KrivovG.G., ShapovalovM.V. & DunbrackR.L. Jr. (2009). Improved prediction of protein side-chain conformations with SCWRL4. Proteins.
12. PonderJ.W. & CaseD.A. (2003). Force fields for protein simulations. Adv. Prot. Chem. 66, 27-85.
13. HessB, KutznerC, Van Der SpoelD, LindahlE (2008). GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. J Chem Theory Comput 4(2): 435. doi:10.1021/ct700301q.
14. ZhouH. & ZhouY. (2002). Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Science, 11, 2714-2726.
15. SamudralaR, MoultJ. (1998). An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. Journal of Molecular Biology 275:893-914.
16. SvergunD.I., BarberatoC. & KochM.H.J. J. (1995) CRYSOL - a Program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. Appl. Cryst. 28, 768-773.

Pushchino

SCF_THREADER with Improved Scoring Function: Generating 3D Protein Models Based on Threading Approach

S.O. Garbuzynskiy, M.Yu. Lobanov and A.V. Finkelstein
Institute of Protein Research RAS
afinkel@vega.protres.ru

For CASP9, we made a server (SCF_THR09) based on our method SCF_THREADER[1] with an improved scoring function[2].

Methods

Selection of templates and target-template alignments for CASP9 targets were done by an extended version of our program SCF_THREADER[1] with a scoring function described in[2-4], which takes into account the following factors:

(1) similarity of sequences calculated by similarity matrices GONNET and BLOSUM50 (this combination was shown[2] to be optimal);

(2) coincidence of secondary structures of the target secondary structure, predicted by PSIPRED[5], and template secondary structure, calculated by DSSP[6];

(3) Miyazawa-Jernigan[7-8] interaction energy of the aligned target residues with the template residues.

If the non-aligned part of the target exceeds 50 amino acids, we treated this non-aligned part separately in the same way as the whole sequence.

Structures of small non-aligned regions within the aligned target region ("loops"), were added based on loops of the same size (and similar conformation of both ends) observed in PDB; the procedure is similar to the "DGLOOP" utility of WHATIF[9].

Availability

The web-server SCF_THR09 is available at casp@chuk.protres.ru

This work was supported by the grants from HHMI (#55005607), MCB (#01200957492) and "Leading Scientific Schools" (#NSh-2791.2008.4) programs, RFBR (#10-04-00162-a), FASI (#02.740.11.0295) and the Dynasty Foundation.

1. 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**, 494-501.
2. Lobanov,M.Y. & Finkelstein,A.V. (2010). Analogy-based protein structure prediction: III. Optimizing the combination of the substitution matrix and pseudopotentials used to align protein sequences with spatial structures. *Mol. Biol. (Moscow)*, **44**, 109-118.
3. Lobanov,M.Y., Bogatyreva,N.S., Ivankov,D.N. & Finkelstein,A.V. (2009). Analogy-based protein structure prediction: I. A new database of spatially similar and dissimilar structures of protein domains for testing and optimizing prediction methods. *Mol. Biol. (Moscow)*, **43**, 665-676.
4. Lobanov,M.Y. & Finkelstein,A.V. (2009). Analogy-based protein structure prediction: II. Testing of substitution matrices and pseudopotentials used to align protein sequences with spatial structures. *Mol. Biol. (Moscow)*, **43**, 677-684.

5. Jones,D.T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195-202.
6. Cabsch,W. & Sander,C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers,* **22**, 2577-2637.
7. Miyazawa,S. & Jernigan,R.L. (1996). Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* **256**, 623-644.
8. Miyazawa,S. & Jernigan,R.L. (1999). Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues. *Proteins,* **34**, 49-68.
9. Vriend,G. (1990). WHAT IF: A molecular modeling and drug design program. *J. Mol. Graph.* **8**, 52-56.

QMEAN
QMEANclust
QMEANdist
QMEANfamily

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

P. Benkert[1,2], M. Biasini[1,2] and T. Schwede[1,2]
[1] - *Swiss Institute of Bioinformatics, Basel, Switzerland*
[2] – *Biozentrum, University of Basel, Switzerland*
pascal.benkert@unibas.ch

We participated in the quality assessment category of CASP9 with four servers. Three servers operate on single models, namely the composite scoring function QMEAN[1,2], QMEANfamily and the new method QMEANdist. The fourth server, QMEANclust[2], is a consensus method which uses QMEAN to prioritise the models.

Methods

As described elsewhere in more detail[1,2], the QMEAN scoring function combines four descriptors based on potentials of mean force with two agreement terms: interaction potentials on Cβ (i.e. residue-level) and all atoms assess long-range interactions; a torsion angle potential over three consecutive amino acids analyses the local backbone geometry of the model and a solvation potential describes the burial status of the residues. The two agreement terms look at the overlap of the predicted and observed secondary structure and solvent accessibility. However, like most of the existing scoring functions, QMEAN has been optimized to assign a relative quality measure to rank the models during the model building process with limited ability to put the quality of the models on a global scale. Still, absolute quality measures are of high importance as they eventually dictate the models usefulness to answer the biological question at hand. We have extended the QMEAN scoring function to produce absolute scores. To correct for size dependency of the scoring function, the model quality estimates were first normalised with respect to the number of interactions or residues in the model. Then, QMEAN Z-scores are calculated which express the likelihood that a model is of comparable quality than an ensemble of high-resolution experimental structure of similar size (manuscript submitted).

QMEANfamily and QMEANdist are additionally enriched with information from evolutionary related proteins. QMEANfamily creates an ensemble of supplementary models for protein sequences sharing at least 40% sequence identity to the target using the starting model as template. The QMEANfamily score is the average QMEAN score of these models covering the protein family. QMEANdist adds an additional term to the QMEAN scoring function that calculates the agreement with residue-level distance constraints extracted from related protein structures to assess the quality of models[3].

In QMEANdist, a pair of C-alpha atoms in the model is scored by comparing the euclidian distance between the atoms to a propensity function calculated from distances observed in related protein structures (templates). The propensity function is essentially a weighted sum of Gaussians, each Gaussian representing one observed distance. The sigma is chosen proportionally to the root mean displacement of the atoms, making the Gaussians wider or narrower for flexible and rigid parts, respectively. The weight of the Gaussian is calculated from the evolutionary distance between target and template. The resulting propensity distribution exhibits small entropy for C-alpha pairs where the templates agree, whereas the entropy gets larger for regions where the template structures deviate. The global score of the model is

then calculated by summing up the individual pairwise scores. The uneven distribution of structures in the PDB leads to clusters of structure with high mutual sequence identity. On one hand, including all of these structures would draw the score towards the large clusters, on the other hand, in absence of any additional information, it is difficult and even questionable, to select a representative. By treating these clusters as units, and downweighting individual structures accordingly, we obtain a detailed statistical description of the cluster.

QMEANclust[2] combines structural density information of by the ensemble of models with the QMEAN scoring function. Compared to CASP8, QMEAN was not used as a pre-filter but rather as a weighting factor in the subsequent consensus calculation. The consensus score of a given model is its weighted mean GDT_TS deviation to all models in the subset.

Results and Discussion

The normalized QMEAN scores gave a slight performance improvement over non-normalized QMEAN for the global correlation. Of the two methods using information from evolutionary related proteins, only QMEANdist is able to significantly boost the performance. For easy and medium targets, the template agreement term is most discriminative for ranking models. For harder targets, the templates become less reliable and additional scores are required. Here the terms of the classical QMEAN help to stabilize the correlation coefficient. For the selection of the best template, however the combination of QMEAN and the template agreement term produces the best results.

The use of consensus information from the ensemble of models (QMEANclust) produces the best global correlation over all targets. However, consensus methods have a limited applicability outside of CASP where typically only a few models are available. The use of structural information from templates has the potential to overtake this role. Results on CASP8 (Table 1) and preliminary results on CASP9 indicate that QMEANdist almost reaches the performance of clustering methods in model ranking and selection.

Table 1. Performance of different QMEAN-based scoring functions on server models of CASP8.

Scoring function	#targets	mean(r)	global r	delta_GDT
QMEAN	121	0.73	0.79	-0.082
QMEANfamily	121	0.74	0.76	-0.089
QMEANdist (template agreement)	120	0.86	0.81	-0.090
QMEANdist (with QMEAN)	121	0.89	0.85	-0.060
QMEANclust	122	0.93	0.93	-0.052

Implementation and Availability

All QMEAN versions have been implemented based on OpenStructure[4] (www.openstructure.org). QMEAN is available on the QMEAN server (http://swissmodel.expasy.org/qmean)[5]. A stand-alone version is available on request.

1. Benkert,P., Tosatto,S.C.E. & Schomburg,D. (2008). QMEAN: A comprehensive scoring function for model quality assessment. Proteins 71, 261-277.
2. Benkert,P., Tosatto,S.C.E. & Schwede,T. (2009). Global and local model quality estimation at CASP8 using the scoring functions QMEAN and QMEANclust. Proteins. 2009;77 Suppl 9:173-80.
3. Paluszewski,M. & Karplus,K. (2009). Model quality assessment using distance constraints from

alignments. Proteins, 15;75(3):540-9.
4. Biasini,M., Mariani,V., Haas,J., Scheuber,S., Schenk,A.D., Schwede,T. & Philippsen,A. (2010) OpenStructure: A flexible software framework for computational structural biology, Bioinformatics, doi:10.1093/bioinformatics/btq481
5. Benkert,P., Künzli,M. & Schwede,T. (2009) QMEAN server for protein model quality estimation., Nucleic acids research, 37, W510-514.

QUARK

See Zhang, Zhang-server

RaptorX

Multiple-template and fragment-free approach to protein modeling

Jinbo Xu[1], Jian Peng[1] and Feng Zhao[1]
[1] –*Toyota Technological Institute at Chicago*
j3xu@ttic.edu

RaptorX is totally different from our previous structure prediction program RAPTOR, which is a pure protein threading program emphasizing on the optimization of a given threading scoring function (by linear programming). By contrast, RaptorX consists of the following major components: pairwise sequence-template alignment[1,2], alignment quality assessment, multiple sequence-template alignment, and fragment-free approach to template-free modeling[3,4]. Even the pairwise sequence-template alignment method in RaptorX is totally different from RAPTOR. RaptorX emphasizes on the design of a good alignment model while RAPTOR on the solution of a given alignment model. RaptorX also pays more attention to protein threading of low-homology proteins. A protein is low-homology if it does not have a lot of non-redundant homologs in the protein sequence databases such as the NR database.

Methods

Pairwise sequence-template alignment. Homologous information has proved to be very powerful in detecting remote homologs and generating accurate alignments, as demonstrated by the excellent profile-based method HHpred. However, profile-based methods do not fare well when proteins under consideration are low-homology. A sequence profile for a low-homology protein, either represented by an HMM or a position-specific scoring matrix, is not good enough to link this protein to its remote homologs. We have developed a profile-entropy dependent scoring function for low-homology protein threading. Our method takes into consideration the number of non-redundant homologs available for the sequence and template and also the sophisticated correlation among various protein features. The relative importance of sequence and structure information is determined by the entropy of a sequence profile. A low-homology usually has a sequence profile with small entropy. When proteins under consideration are low-homology, our threading scoring function will rely more on structure information; otherwise, sequence profile.

Alignment quality assessment. We developed a method that can predict the global and absolute quality of a sequence-template alignment, which is defined as the quality (measured by GDT-TS/TMscore) of the 3D model built from this alignment by MODELLER (with default parameters). Our method does not need to actually build such a 3D model in order to do alignment quality assessment since our method does not use any 3D structure information. Instead, our method uses only information extracted from an alignment. We have tested our method on several large datasets including data generated by RAPTOR for previous CASP (Critical Assessment of Structure Prediction) events. The MAEs (mean of absolute errors) of both the predicted TM-score are ~0.05 and the Pearson correlation coefficient (PCC) between the predicted GDT-TS/TM-score and the real is ~0.95. We used this method in CASP9 for template selection.

. **Multiple sequence-template alignment (RaptorX-MSA).** Based upon the pairwise sequence-template alignment, we have developed a new probabilistic method to align a single target sequence to all of its top templates. The multiple sequence-template alignment is then fed into MODELLER to generate a 3D model. The distinguished feature of RaptorX-MSA is that it not only can increase coverage for the target (by copying structures from multiple templates), but also improve pairwise sequence-template alignment by using distance constraints in multiple templates. Existing multiple-template methods either

simply assemble pairwise alignments into a single multiple alignment or use existing multiple sequence alignment tools such as T-Coffee, MUSCLE and ProbCons to generate a multiple alignment. Therefore, these methods usually cannot correct alignment errors in the pairwise alignments. That is, the errors in a pairwise alignment will persist in the multiple-alignment. By contrast, our method can correct errors in a pairwise alignment by using structure information in multiple templates and thus, improve the final model quality. In addition, the multiple sequence alignment tools such as T-Coffee, MUSCLE ProbCons, and MAFFT are not very good at the alignment of remote homologs, especially when the proteins are low-homology.

Fragment-free approach to template-free modeling (RaptorX-FM). The popular fragment assembly method generates conformations by restricting the local conformations of a protein to short structural fragments in the PDB. This method may limit conformations to a subspace to which the native fold does not belong because a protein with a new fold may contain some structural fragments not in the PDB. RaptorX-FM is a probabilistic method that can sample conformations in a continuous space without using any structure fragments to assemble a protein conformation. Therefore, RaptorX-FM can be used to predict structures for the targets with a truly new fold. In addition, we also used RaptorX-FM in CASP9 to fold the unaligned regions in the two ends of a target.

RaptorX-Boost/RaptorX. RaptorX-Boost uses TASSER to generate a 3D model from the alignment generated by RaptorX-MSA. Looks like that RaptorX-Boost performs worse than RaptorX-MSA in CASP9. Maybe it is because we did not use TASSER in a correct way. RaptorX is a combination of RaptorX-MSA, RaptorX-Boost and RaptorX-FM. When the target appears to be easy (i.e., predicted GDT-TS at least 80), RaptorX used the results from RaptorX-MSA. When the target appears to be hard, RaptorX used the results from RaptorX-Boost. When no reliable templates can be identified for a target (i.e., predicted GDT-TS less than 40), RaptorX used RaptorX-FM to generate five models.

Results

The unofficial CASP9 evaluation by Dr. Zhang at the University of Michigan indicates that RaptorX/RaptorX-MSA indeed outperforms our previous program RAPTOR significantly. Compared to RAPTOR, RaptorX improves alignment accuracy significantly, but template selection is still a major issue with RaptorX.

We have tested our fragment-free approach to template-free modeling (i.e., RaptorX-FM) using some targets in previous CASP events. RaptorX-FM performs very well on mainly alpha proteins and small beta proteins.

Our low-homology threading method works well on the two public benchmarks SALIGN and ProSup. We use TM-score to evaluate the reference-independent alignment accuracy of the alignments generated by our method and HHpred for the protein pairs in these two benchmarks. The alignments generated by our method in total have TM-score 66.77 and 132.85 on Prosup and SALIGN, respectively. By contrast, HHpred achieves TM-score 56.44 and 119.83 on Prosup and SALIGN, respectively. Our method is better than HHpred by 18.3 and 10.9% on ProSup and SALIGN, respectively. A Student's *t*-test indicates that our method excels HHpred with *P*-values being $3.77E{-}11$ and $9.83E{-}13$, respectively. The unofficial CASP9 evaluation by Dr. Zhang further confirms that our threading method outperforms HHpred on hard targets, many of which are low-homology proteins.

Availability

RaptorX temporarily is available at http://velociraptor.ttic.edu. We are testing the server more extensively, writing documents and will move it to a new machine.

1. Jinbo Xu. Boosting protein threading accuracy. In the Proceedings of the 13th International Conference on Research in Computational Molecular Biology (RECOMB), Lecture Notes in Computer Science, Vol. 5541, pp. 31-45, 2009. Springer.
2. Jian Peng and Jinbo Xu. Low-homology protein threading. Bioinformatics (Proceedings of ISMB 2010), 2010.
3. Feng Zhao, Jian Peng and Jinbo Xu. Fragment-free Approach to Protein Folding Using Conditional Neural Fields. Bioinformatics (Proceedings of ISMB 2010), 2010.
4. Feng Zhao, Jian Peng, Joe DeBartolo, Karl F. Freed, Tobin R. Sosnick and Jinbo Xu. A probabilistic and continuous model of protein conformational space for template-free modeling. Journal of Computational Biology, 2010.

RBO-PROTEUS

Combining Model-based Search with a Balanced Exploration-Exploitation Template Search

N. Mahmood[1], TJ Brunette[2] and O. Brock[1]
[1] - *Robotics and Biology Laboratory, School of Electrical Engineering and Computer Science, Technische Universität Berlin, Einsteinufer 17-EN 10, 10587 Berlin, Germany,* [2] - *Department of Biochemistry, University of Washington, Seattle, WA 98195, USA*
nasir.mahmood@tu-berlin.de

De novo protein structure prediction involves an extensive search in a high-dimensional conformational space. Frequently used Monte Carlo methods are memory-less and often rely on random walks for exploration of space, ignoring the possibility of search guidance towards meaningful regions identified by information from previous steps. We present a novel search technique by combining model-based search[1] (MBS) with balanced exploration-exploitation template search (BEETS). MBS builds an approximate model of underlying function and incrementally refines that model as search progresses. Depending upon state of the model, BEETS decides exploitation level of structural information in Protein Data Bank (PDB).

Methods

Model-based search can be viewed as an active learning technique[2] that treats the information obtained during search as a valuable insight[3] and exploits that information to guide further search space exploration in a highly efficient manner. We refer to the learnt representation of relevant regions of energy landscape as a model. Initially, this model is empty or coarse but gradually becomes more informative for succeeding move steps. The acquisition of high quality information depends upon three algorithmic features: 1) characterization of regions as funnels, 2) assessment of relevant funnels by determining whether all samples in a region share biological characteristics, and 3) distribution and coordination of computational resources in accordance with assessment.

There have been two forms of exploitation of PDB structural information: 1) very small fragments used by fragment assembly methods, and 2) large portions of protein structures to build models by homology modeling methods. Both fragment assembly and homology modeling have limitations. BEETS, our novel extension to MBS, is able to use structural information from the PDB of any length—starting with fragments all the way up to large portions of proteins. The key advancement is a method to balance exploration and exploitation: when BEETS selects a chunk of structure from the PDB to guide search, it cannot be sure that this is a good decision. But BEETS can recognize and recover from mistakes As a result, the conformational space search consisting of MBS and BEETS can guide search more effectively by using a wider range of structural information more effectively than previous search methods.

Availability

An implementation of MBS and BEETS can be obtained by contacting the authors.

1. Brunette, TJ & Brock, O. (2008). Guiding conformation space search with an all-atom energy potential. *Proteins* **73**, 958-972.
2. Cohn, D.A., Ghahramani, Z. & Jordan, M.I. (1996). Active learning with statistical methods. *J. Art. Intell. Res.* **4**, 129–145.
3. Glover, F. & Laguna, F. (1997). Tabu Search. Kluwer Academic Publisher.

RecombineIt

Fully automated modeling server based on scoring of models by MQAPmulti and recombination of best-scoring fragments.

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

Tertiary structure predictions (TS)

or the prediction of the 3D structures in CASP9, we have developed a fully automatic procedure comprising collection of 3rd-party models, local assessment of model quality by MQAPmulti developed in our laboratory, and recombination of best-scoring fragments. The whole procedure comprises the following five steps:

1. The method collects models for a given target sequence. In CASP9 we download all server models.
2. Each model is scored by MQAPmulti (see another abstract in the book). Both local residue deviations and global model scores are predicted. Five models with the best global score are selected.
3. All input models are divided into partially overlapping fragments containing 1 or 2 secondary structure elements, depending on the target size.
4. All possible combinations of fragments are ranked (without explicitly generating 3D models for each combination). To rank a given combination of fragments, the sum of local MQAPmulti scores is calculated for all residues. In addition, a complex penalty system is applied. Penalty is given for fragments:

a) derived from models with different folds; b) derived from models with folds different from the folds of top 5 models selected in step 2; c) if the area of overlap between fragments exhibits different structure.

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

Refinement

Different technique was tested on the refinement targets. In the refinement category the starting models were refined using REFINER program. Then, according to REFINER^2 scoring function, 1000 best decoys were reranked by MQAPmulti program. The highest-ranking assembled model reported as the final result.

1. Sali, A. & Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **234**, 779-815.
2. Boniecki,M., Rotkiewicz,P., Skolnick,J., Kolinski,A. (2003). Protein fragment reconstruction using various modeling techniques. *J. Computer Aided Molecular Design* **17**, 725-737

SAM-T2k-server
SAM-T06-server
SAM-T08-server

Old servers serve as historical baseline for evaluating progress in prediction methods

Kevin Karplus
University of California, Santa Cruz
karplus@soe.ucsc.edu

Due to two years without funding, the group at UCSC did not test any new methods in CASP9, but only provided three old servers (SAM_T02, SAM_T06, and SAM_T08) for historical comparisons. Note that the SAM_T08 server is essentially the same as the SAM_T08_server in CASP8, not even including improvements using in SAM_T08_human predictions in CASP8.

The databases searched by these three servers have been kept up to date, but the methods have not changed (except in small unavoidable ways as some software was broken by changes to the server compilers or operating systems, and had to be replaced by newer versions that would run on the current system). Thus the servers represent a fairly stable baseline for judging the difficulty of targets, which is important for determining whether there has been progress in other methods.

Unless we get some funding for protein structure prediction in the next year, we will not be participating in future CASP experiments at all, and will not be doing further work in protein structure prediction.

SAMUDRALA

Automated Model Refinement Using Knowledge Based Constraints Consensus Refinement Methods

Michael Zhou, Jeremy Horst, Raymond Zhang, Ram Samudrala
Computational Biology Group, University of Washington
{mzhou,jhorst,raymond,ram}@compbio.washington.edu

Selection of models used in refinement

All starting models for our model refinement methods were chosen from the CASP9 released server models. Initial models were either taken as the closest all-atom RMSD models from the given starting model for refinement targets, or from the I-TASSER[i] server models for tertiary structure targets.

Extraction of a consensus of constraints from initial models

Refinement is achieved by taking consensus constraint sets as derived from the initial models and rebuilding the models from these constraints using CYANA[ii] restrained torsion angle dynamics simulations. The success of this method relies on being able to identify sets of constraints that are both accurate and plentiful. From this initial set of constraints, the distances are weighted using our knowledge based residue specific all atom probability discriminatory function (RAPDF) and compiled to refine the constraint sets.

The distances selected for inclusion in constraint sets are one in which at least four of the five models show consensus within 0.5 Ångstrom distance bins. These distances are scored with RAPDF, and compiled using batches of constraints, starting with the best consensus distances by RAPDF, resulting in one distance for each pair of residues. We then use RAPDF and the amount of consensus (ie. Consensus among four or five models) to create three constraints sets (12 Å, 16 Å, 20 Å).

Rebuilding models from constraints and model selection

The constraints sets are used in fifty rounds of CYANA simulations, using Ramachandran plots to derive probabilities for torsion angles. A total of three thousand conformations are generated, with each of the fifty rounds producing twenty models. All models are then minimized by ENCAD[iii], and side chains are optimized by SCWRL 3.0[iv]. Lastly these final models are scored using RAPDF and the top five models are selected for submission to CASP.

Submission as a human predictor

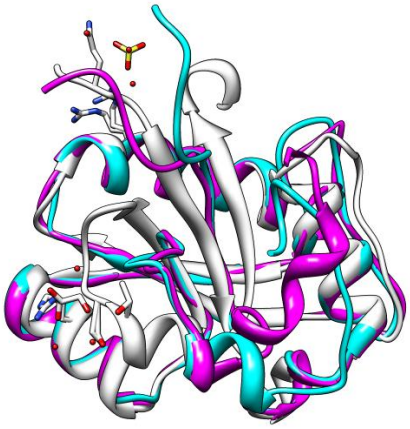
Although the entire process was automated, the time required for CYANA simulations to test the torsional space prevents us from submitting as a "server" predictor. However, a server for use to the public is nearing completion and will be hosted at http://protinfo.compbio.washington.edu/refine.

Technical difficulties with some submissions

Due to technical issues, refinement models up to TR574 were not able to be sent through the refinement method as described. Instead, identified problem loops were rebuilt using the mcgen_exhaustive_loop and mcgen_semfold_loop functions of the RAMP suite[v]. These models then

were minimized by ENCAD, and SCWRL 3.0. Lastly, the best five models by RAPDF score were submitted. In addition, for certain tertiary structure models we were not able to derive consensus sets with sufficient robustness to significantly improve the initial models. In these cases, one of the best models selected by RAPDF was submitted instead.

Refinement target TR606



For this target, our submitted model (purple) is a refined model of the initial given model (cyan). This represents a CαRMSD improvement of 0.179 from an initial model with a 6.127 CαRMSD.

SAMUDRALA

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

B. Buttrick, A.A. Laurenzi, J.A. Horst and R. Samudrala
Computational Biology Group, University of Washington
ram@compbio.washington.edu

Each atom in a given protein provides a quantifiable contribution to the overall function of the protein. The degree of functional importance of atoms in a residue and corresponding positions can be thought of as the "functional signature" of a protein. We have developed a combination of knowledge-based techniques to determine the functional importance of each residue and corresponding position to elucidate the structural and the functional interplay of individual residues and positions. The techniques yield a meta-functional signature (MFS), a collection of continuous values representing the functional significance of each residue in a protein. MFS values were calculated for each target in the CASP9 experiment for the blind prediction of protein functional sites.

Methods

Meta-functional signature calculation (MFS)

The sequence-based protein meta-functional signatures were calculated using sequence conservation, evolutionary conservation, and amino acid type score (MFS1)[5]. Briefly, the sequence conservation score is calculated from positional relative entropy using amino acid frequencies estimated by a hidden Markov model; the evolutionary conservation score was calculated by a state to step ratio of residue type changes in a phylogenetic tree built for each position; and the amino acid type score was derived from the prior probability of an amino acid being identified as functionally important in two databases of catalytic and ligand binding residues. MFS scores were visually represented using UCSF Chimera by shading each residue according to its MFS score. Observation of spatial clustering of high-scoring residues added confidence to the MFS predictions.

The second generation of the meta-functional signature (MFS2)[1] builds upon MFS1 by including sequence-derived predictions of secondary structure, level of solvent exposure, disorder, disulfide bonds, domain breaks, and nonlocal contacts. Concordant function of residues within 5 positions is predicted based on the expected secondary structure. The stability of each residue is predicted using the amino acid type count, the mean and distribution of sequence conservation scores, the probability of nonlocal contacts and the number of contacts within a concentric shell. All of these structural features are predicted using the suite of software from the Jianlin Cheng group.

When a calcium ion was identified as the ligand for a target sequence, a version of MFS2 that was trained by logistic regression for the specific function of calcium binding was used (MFS2Ca)[2]. MFS2Ca was trained on a database of high-resolution calcium binding chains with less than 60% sequence identity. The sensitivity of MFS to metal ion binding is increased because fewer residues bind to metal ions, and the bonds necessary for coordinating metal ions in functional sites generally arise from the same types of residues involved in catalytic functionality (sought by MFS), such as histidine, cysteine, aspartic and glutamic acid.

Inclusion of homologous structures

In cases where ligand-bound determined protein structures showing homology to the target sequence (templates) were available, this information was used to aid in functional site prediction. To incorporate homology information in functional site prediction, HHpred[3] was used to search the target

sequence against the PDB to identify templates. We developed a tool that takes as input a set of templates and aligns them to predicted models using the matchmaker function of UCSF Chimera[4], effectively mapping the ligand onto the predicted models. An 'agreement value' is calculated for each template-model alignment representing how well the MFS predictions align with those suggested by homology to the ligand-bound template. The agreement value is calculated by totaling the MFS scores of each residue from the model within 0.5 Å plus the sum of the Van der Waals distances of the mapped ligand and dividing the total by the number of residues considered yielding a value between 0 and 1. Model-template alignments with the highest agreement values were examined using Chimera with the MFS scores represented visually. Alignments showing high MFS scores clustered around the mapped ligand were chosen as functionally important.

Summary and conclusions

Our approach to predict functional sites is powerful because we use two independent measurements of functional significance, one derived solely from sequence information (MFS) and another from homology to determined protein structures. In cases where homologous ligand-bound structures were available, predictions derived from MFS and homologous structures converged adding confidence to our predictions. Therefore, the use of sequence-derived information, specific training of MFS for known functions (MFS2Ca), and the use of structural information from alignment of predicted structures to homologous ligand-bound structures enhances sensitivity and precision of functional site prediction.

Availability

MFS1 is available here: http://protinfo.compbio.washington.edu/mfs/. MFS2 and MFSCa will soon be available at that URL as well.

1. Horst,J.A., Laurenzi,A.A., Buttrick,B., Zhou,M., & Samudrala,R. A generalized approach to active site prediction by reassembling specific protein meta-functional signatures (to be submitted).
2. Horst,J., Samudrala,R. (2010) A protein sequence meta-functional signature for calcium binding residue prediction. *Pattern Recognition Letters.* **31**, 2103-2112.
3. Söding,J., Biegert,A., & Lupas,A.N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research.* **33**, W244--W248 (Web Server issue).
4. Pettersen, E. F., Goddard, T. D., Huang,C.C, Couch,G.S., Greenblatt,D.M., Meng,E.C. & Ferrin,T.E. (2004). UCSF Chimera - a visualization system for exploratory research and analysis. *J Comput Chem.* **25(13)**, 1605-1612.
5. Wang,K., Horst,J.A., Cheng,G., Nickle,D.C., & Samudrala,R. (2008). Protein Meta-Functional Signatures from Combining Sequence, Structure, Evolution, and Amino Acid Property Information. *PLoS Comput Biol.* **4(9)**, e1000181.

sbtJ

Kentaro Tomii
National Institute of Advanced Industrial Science and Technology (AIST)

We used our automated system of profile–profile comparison method, called FORTE[1,2], to obtain the set of target–template alignments for each target. The FORTE system utilizes position-specific score matrices (PSSMs) of both the target and templates to build a sequence-structure alignment and predict the protein structure of target sequence. To obtain an optimal alignment of a target sequence profile onto a template profile, we employ the global–local algorithm which is based on the global alignment algorithm with no penalty for the terminal gaps. The statistical significance of each alignment score is estimated by calculating Z-scores. For long loop regions we used the alignments between a target protein and many homologous proteins of templates, and local alignments between a target protein and other templates. Then, based on those alignments, we constructed and exhaustively evaluated 3D models with MODELLER. The 10 models for each alignment of top hits were constructed. Candidates among those models were selected using quality scores.

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., and Motono, C. (2005). Protein structure prediction using a variety of profile libraries and 3D verification. Proteins, 61, 114-21.

Scheraga

Protein-structure prediction with physics-based UNRES force field using multiplexed replica exchange molecular dynamics

Yi He[1], Adam Liwo[1,2], Stanisław Ołdziej[1,3], Cezary Czaplewski[1,2] and Harold A. Scheraga[1*]
[1] – Baker Laboratory of Chemistry, Cornell University, Ithaca, NY 14853-1301, [2] – Faculty of Chemistry, University of Gdańsk, Sobieskiego 18, 80-952 Gdańsk, Poland,
[3] – Intercollegiate Faculty of Biotechnology, University of Gdańsk, Medical University of Gdańsk, Kładki 24, 80-822 Gdańsk, Poland,
has5@cornell.edu

The structures of the target proteins were predicted by a procedure which consists of the following three steps. First, UNRES was employed to carry out Multiplexed Replica Exchange Molecular Dynamics (MREMD)[1] for the target proteins. Second, based on MREMD simulation results, Weighted-Histogram Analysis Method (WHAM) analysis was used to calculate the relative free energy of each structure of the last slice of the MREMD simulation. Third, cluster analysis was employed to cluster the structures from the MREMD simulation. Five clusters with the lowest free energy were chosen as final submitting candidates in most cases.

In the UNRES model, a polypeptide chain is represented by a sequence of α-carbon atoms connected by virtual bonds with attached side chains. Two interaction sites are used to represent each amino acid: the united peptide group (p) located in the middle between two consecutive α-carbon atoms and the united side chain (SC). 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. The cumulant expansion enabled us to determine the functional forms of the multibody terms in UNRES. Recently, temperature dependence has been introduced to UNRES, the energy function was reparameterized and its reliability has been shown on test proteins. [2,3]

To obtain better sampling of the conformational space in the UNRES model, we used MREMD. MREMD searches were carried out in the range of temperatures from T=250 K to T=500 K. To take advantage of massive-parallel computations to run simulations in real time, we used our recently developed fine-grained UNRES code. To extract the candidate conformations from the results of MREMD simulations, we used a procedure developed in our recent work. First, WHAM analysis was used to calculate free energy of each structure from the last 100 snapshots of each trajectory from MREMD simulations (totally 6400 structures). Then cluster analysis was employed to cluster all the structures used in WHAM analysis. The conformations closest to the average structures corresponding to the found clusters were considered as candidate models. Based on WHAM and cluster analysis results, an average free energy of each cluster was calculated. The clusters were ranked according to increasing free energy. UNRES 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. [4,5] In order to speed up the search for larger proteins, information from secondary structure prediction by PSIPRED[6] was used in the generation of the initial structures.

1. Czaplewski, C., Kalinowski, S., Liwo, A., Scheraga, H.A., (2009) Application of multiplexed replica exchange molecular dynamics to the UNRES force field: tests with alpha and alpha+beta proteins. *J. Chem. Theor. Comput.*, **5**, 627-640.
2. Liwo, A., Khalili, M., Czaplewski, C., Kalinowski, S., Ołdziej, S., Wachucik, K. & Scheraga, H.A. (2007) Modification and optimization of the united-residue (UNRES) potential energy

function for canonical simulations. I. Temperature dependence of the effective energy function and tests of the optimization method with single training proteins. *J. Phys. Chem. B,* **111**, 260-285.
3. He, Y., Xiao Y., Liwo, A., Scheraga, H.A., (2009) Exploring the parameter space of the coarse-grained UNRES force field by random search: selecting a transferable medium-resolution force field. *J. Comput. Chem.*, **30**, 2127-2135.
4. Kazmierkiewicz, R., Liwo, A., Scheraga, H.A.. (2002) Energy-based reconstruction of a protein backbone from its alpha-carbon trace by a Monte Carlo method. *J. Comput. Chem.*, **23**, 715-723
5. Kazmierkiewicz, R., Liwo, A., Scheraga, H.A., (2003) Addition of side chains to a known backbone with defined side-chain centroids. *Biophys. Chem.*, **100**, 261-280.
6. McGuffin, L.J., Bryson, K., Jones, D.T., (2000) The PSIPRED protein structure prediction server. *Bioinformatics* **16**, 404-405.

SchroderLab

Restrained-ensemble physics-based refinement

A. Wojtyczka[1], G.F. Schröder [1]
[1] *Forschungszentrum Jülich,*
gu.schroeder@fz-juelich.de

Our approach is based on physics-based modeling to refine protein structures. We therefore only submitted predictions for the refinement targets. We performed simulated annealing molecular dynamics simulations. However, instead of using just one single structure, we modified the standard simulation setup which allowed us to simulate eight copies of the same protein at the same time and to apply restraints to keep these eight structures close to each other.

The rationale was that the effective energy landscape would be smoother for the ensemble of restrained structures and that cooperativity helps to overcome energy barriers. Furthermore the distribution of conformations becomes sharper, leading to higher precision of the obtained ensemble of refined structures.

For each target we performed 1000 short (100ps) simulated annealing runs in explicit water using Gromacs with the Amber03 force field. The best five structures were picked by clustering based on nearest neighbor counts. This clustering is far from optimal and severely limits the quality of the obtained candidate structures; we are selecting candidate structures almost at random from the ensemble of refined structures.

Seok
Seok-server

Prediction of Ligand-binding Sites by Molecular Docking on Protein Tertiary Structure Models

Lim Heo, Woonghee Shin and Chaok Seok*
Department of Chemistry, Seoul National University, Seoul, Republic of Korea
chaok@snu.ac.kr

Knowledge on possible ligand-binding sites of a protein is valuable for understanding its function and designing novel molecules that regulate the function. In CASP9, we demonstrate that molecular docking simulations with carefully designed energy functions can be successfully applied to ligand-binding site prediction.

Methods

For both server and human binding-site predictions, we predicted ligand-contacting residues of a protein from the predicted docking pose of putative ligands after molecular docking simulations. Identities of ligands that can possibly bind were predicted by scoring ligands in the experimental structures of homologous proteins found using HHsearch[1]. The scoring function used for ligand prediction evaluates each ligand considering the number of homologous proteins that contains the ligand, similarities of the homologous proteins to the target protein, and the degree of conservation of the binding positions of the ligand in the homologous proteins. The best scoring ligands were docked on the protein tertiary structure model generated by our own server in the server binding-site prediction and on additional models submitted by several other servers in the human binding-site prediction.

New docking scoring functions and sampling methods recently developed were employed for molecular docking (manuscript in preparation). For non-metal ligands, conformational space annealing was applied to the optimization of a docking energy function expressed as a linear combination of the AutoDock[2] energy and additional energy terms extracted from the protein-ligand interactions observed in the experimental structures of the homologous proteins. For metal ligands, Monte Carlo simulations were carried out with a docking energy that consists of electrostatic energy, penalty function for atomic clash, orientation-dependent metal coordination energy, and distance and angle energy derived from the homologous proteins. The docking poses obtained by each simulation were clustered, and the largest cluster was selected as the final docking pose. The ligand-contacting residues were selected from the docking pose by the standard contact criterion used in CASP. The whole procedure is fully automatic. Our method not only predicts contact residues but also provides putative binding poses and atomic details of protein-ligand interactions that can be valuable for further applications.

For human predictions, additional efforts were exerted on ligand prediction, protein structure model selection, and contact residue selection as follows. In the ligand prediction step, biological databases were searched to obtain information on function. Protein structure models submitted by several servers were used in docking simulations and the final protein model with docked ligands was selected after consideration of physicochemical aspects of the protein-ligand binding. Contact residues were determined after visual inspection of side chain orientations, especially for metal-containing targets.

Results

We submitted server predictions for all CASP9 FN targets and human predictions for 69 CASP9 FN targets. Prediction results for the 13 targets with biologically meaningful ligands in the experimental structures released by Sep. 20, 2010 were analyzed. Our server achieved the average accuracy of 78.3%, coverage of 69.0%, and MCC-score of 0.715 for these targets. Our human method showed further

improvement over the server results, giving the average accuracy of 82.1%, coverage of 73.3%, and MCC-score of 0.764.

Availability

1. Söding,J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics*. **21**, 951-960.
2. Morris,G.M., Goodsell,D.S., Halliday,R.S., Huey,R., Hart,W.E., Belew,R.K. & Olson,A.J. (1988). Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem*. **19**, 1639-1662.

Seok
Seok-server

Template-based Protein Model-building and Refinement of Unreliable Local Regions by Global Optimization

Junsu Ko, Hahnbeom Park, Lim Heo, Juyong Lee, Hasup Lee and Chaok Seok*
Department of Chemistry, Seoul National University, Seoul, Republic of Korea
chaok@snu.ac.kr

Predicting protein tertiary structures with accuracies beyond the limit available from the best single templates is one of the major challenges in current template-based modeling. In CASP9, we tackled this problem by applying model building and model refinement methods that employ global energy optimization.

Methods

For each target protein, multiple templates were selected by re-ranking the homologous proteins detected by HHsearch[1], and a multiple sequence alignment (MSA) of the selected templates and the target was generated by PROMALS3D[2]. The new scoring function for the re-ranking and template selection that was trained on the previous CASP targets was applied. Template-based models were then built by MODELLER-CSA[3] which optimizes the MODELLER energy derived from a given MSA using conformational space annealing. Side chains were subsequently re-modeled. In the refinement stage, unreliable local regions (ULRs) of the models were predicted and re-modeled by a recently developed refinement protocol (manuscripts in preparation). ULRs are often found in the regions for which no proper template information exists, such as loops or terminals. ULRs were predicted by a newly developed model-consensus method. The ULR refinement method searches the conformational space using fragment assembly, analytical loop closure[4], and conformational space annealing. The energy functions for ULR optimization consist of physics-based energy terms and knowledge-based potentials and were trained on separate non-redundant training sets.

For human predictions, the template selection procedure was supplemented by visual inspection, and MSAs were adjusted manually. The same automatic protocol for ULR refinement was applied.

Results

The experimental structures released by Sep. 20, 2010 were analyzed to evaluate the effectiveness of our template selection and model refinement methods. Our template selection scheme was effective especially for hard TBM targets, giving higher TM-scores for the top-ranking templates than those for the best templates given by HHsearch. The refinement stage further improved the model qualities consistently. Significant differences between the model structures before and after refinement were found when TM-scores were compared. Model qualities for the re-modeled local regions were also enhanced, with decreased local RMSD for 75% of the re-modeled regions. These two new components contributed to the high performance of our protein structure prediction protocol over the whole range of TBM targets. Overall, Seok-server occupies the 4th position among the server methods when ranked by the sum of TM-score and HB-score for the first models, according to the assessment by Zhang and co-workers[5].

Availability

A web server for protein loop modeling is under construction at http://loop.neosgen.net. Web service for the full protein structure prediction is also in preparation.

1. Söding,J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics.* **21**, 951-960.
2. Pei,J., Kim,B. & Grishin,N.V. (2008). PROMALS3D: a tool for multiple sequence and structure alignment. *Nucleic Acids Res.* **36**, 2295-2300.
3. Joo.K., Lee,J., Seo,J., Lee,K., Kim,B. & Lee,J. (2009). All-atom chain-building by optimizing MODELLER energy function using conformational space annealing. *Proteins* **75**, 1010-1023.
4. Lee,J., Lee,D., Park,H., Coutsias,E. & Seok,C. (2010). Protein loop modeling by using fragment assembly and analytical loop closure. *Proteins*, in press.
5. http://www.zhanglab.ccmb.med.umich.edu/casp9/.

Sessions

A human lost in the grey zone

R.B. Sessions
University of Bristol, UK.
r.sessions@bris.ac.uk

The author has been building homology (comparative) models for about 20 years to aid the planning and interpretation of experiments for many biochemists. These models have ranged from modest early efforts[1] to more speculative but potentially more interesting models,[2,3] edging into the grey zone of homology modelling (i.e. below 30% sequence identity). These models have typically been informed and refined according to experimental feedback, usually site-directed mutagenesis. Deciding to participate in CASP9 turned out to be something of a "baptism of fire" in that the semi-automatic methods used here to construct even one model for each of the 60 "human targets" neither allowed enough time for much application of "human intuition", nor was the "decision space" small enough for this human to apply anything much more than guesswork.

Methods

Every model is template based. Template structures were identified using BLAST and PSI-BLAST for targets t0515 to t0600, HHpred was used for t600-t640. Sequence alignments were performed using ClustalW2 and human judgement. A simple filter (gf_mutate) was used to convert the sequence alignment into an InsightII macro for residue replacements. The InsightII loop builder was used to build InDels and severe sidechain clashes were relieved with the rotamer tools. Each model was soaked with a 10 Angstrom layer of water and refined with tethered (position-restrained) energy minimization using Discover 2.98.

In those few cases where the bioinformatics tools failed to find a template, members of the set of SCOP sequences of length (target_length-15 to target_length+15 residues) were each aligned with the target using ClustalW2. The alignments were ranked by identity, similarity and gap-penalty and the "best" template chosen by human inspection and consideration of the predicted secondary structure (JPRED).

Results

To be analyzed.

Availability

All the major tools used are freely or commercially available.

1. Kelly,M., Sessions,R.B., Muirhead,H. (1992), A prediction of the tertiary structure of Phospholipase A2 from synovial fluid and a model of substrate binding. Bioorg. Med. Chem. Lett. **2**, 553-558.
2. Wilson,M.C., Meredith,D., Bunnun,C., Sessions,R.B., Halestrap,A.P. (2009), Studies on the DIDS-binding site of monocarboxylate transporter 1 suggest a homology model of the open conformation and a plausible translocation cycle. J. Biol. Chem. **284**, 20011-20021.
3. Kupzig,S., Bouyoucef-Cherchalli,D., Yarwood,S., Sessions,R., Cullen,P.J. (2009), The ability of GAP1(IP4BP) to function as a Rap1 GTPase-activating protein (GAP) requires its Ras GAP-related domain and an arginine finger rather than an asparagine thumb. Mol. Cell. Biol. **29**, 3929-3940.

SHORTLE

Protein structure prediction with statistical potentials and genetic algorithms

D. Shortle
The Johns Hopkins University School of Medicine
Baltimore, MD 21205
gdshortl1@jhmi.edu

Methods

The principal focus of the group is energy-based refinement employing a genetic algorithm driven by several statistical potentials[1,2]. This focus, plus our very limited resources, have constrained us to use the tarball of CASP server models as the input structures for refinement for both TBM and FM targets.

All human/server targets were attempted. One refinement protocol was applied to targets that were straightforward template-based modeling targets, whereas a second protocol was applied to all others. Briefly, the GA program loads the model structures and applies a conventional genetic algorithm to those models that include all residues and CG atoms. Heavy side chain atoms are added, a grid search of major rotamers is conducted at each position, and a list of energy terms is scored. The N best structures (ranked by a simple sum of z-scores for a specified list of parameters) are selected as the initial population. Two parents are picked at random, a recombinant is formed by a single or a double random cross-over(s), and a brief minimization carried out with Monte Carlo moves applied to backbone dihedrals, bond angles, and bond distances. After the population of structures has increased by N recombinants, a survival or fitness selection restores the base population to N. Several fitness selection criteria were used, with some working much better than others.

To reduce the rate of random loss of diversity, the genetic algorithm is run as a series of epochs, consisting of 3 to 5 generations, with the best structures after the final generation saved to file. Multiple cycles through the same epoch accumulates an ensemble of structures with different, more-or-less random subsets of backbone structure retained. Epoch1 consisted of 3 generations with N = 120, initialized by 120 randomly selected structures from the tarball, with scoring for atom solvation, implicit side-chain solvation, and atom-pair interactions; 50 structures are saved after these 3 generations. Epoch2 consisted of 4 generations, N = 200, initialized with randomly selected structures from epoch1. Epoch3 consisted of 5 generations with N = 200. The submitted model had the best sum of z-scores for these parameters.

For targets that were not straightforward TBM, the tarball structures were rebuilt from low backbone energy[2] fragments taken from 6000 high resolution crystal structures, and 1000 of these were used instead of tarball structures themselves. The population size during refinement was 200-500 for all epochs, with much larger pools of partially refined structures accumulated at the intermediate stages.

Results

Self evaluation of 32 targets with released structures and with the highest probability of being TBM targets demonstrated that our MODEL_1 structure was among the most accurate (Ca-RMSD) 5% for half of these targets and in the top 10% in 23 cases. For more than three-quarters of these targets, our MODEL_1 was closer to the experimental structure than MODEL_1 submitted by some of the best servers, including Zhang-Server, RaptorX, and BAKER-ROSETTASERVER.

On the more difficult TBM targets and the NF targets, our strategy worked poorly for reasons we are still trying to understand. Although we attempted the 13 refinement targets, a serious glitch in our protocol lead to the accidental submission of the model to be refined. All we can say is "aaargh".

1. Fang, Q & Shortle, D (2006) Protein refolding *in silico* with atom-based statistical potentials and conformational search using a simple genetic algorithm J. Mol. Biol**. 359**:1456-1467.
2. Fang Q & Shortle, D. (2005) A consistent set of statistical potentials for quantifying local side- chain and backbone interactions. Proteins **60**: 90-96.

SiteHunter

SiteHunterPro: a combined approach for the prediction of functional sites in proteins

M. Brylinski, M. Gao and J. Skolnick
Center for the Study of Systems Biology, School of Biology, Georgia Institute of Technology, 250 14th Street NW, Atlanta, GA 30318
skolnick@gatech.edu

Exhaustive exploration of molecular interactions at the level of complete proteomes requires efficient and reliable computational approaches to protein function inference. Here, we present SiteHunterPro, an automated webserver for the prediction of protein-ligand, protein-metal and protein-DNA interactions based on protein threading. SiteHunterPro comprises three method components: FINDSITE[1] that detects binding pockets for small molecules, FINDSITE-metal[2] that predicts metal-binding sites, and DBD-Threader[3] that identifies DNA-binding sites.

Methods and Results

It is well established that protein threading is capable of detecting remote, yet evolutionary related homologues. Conservation of functional sites among homologous proteins allowed us to develop FINDSITE, a highly accurate method for ligand-binding site prediction and functional annotation. FINDSITE employs template identification, structure superimposition and binding site clustering as follows: First, for a given target sequence, structure templates are selected by three threading procedures: PROSPECTOR_3[4], Sparks2[5] and SP3[6]. Subsequently, template structures bound to ligands are identified and superimposed onto the target protein structure using the structural alignment algorithm fr-TM-align[7]. In CASP9, we used TASSER[8] models as the reference structures. After the superimposition, putative binding sites are inferred through the clustering of the template ligands, and the predicted sites are ranked according to the number of templates that share a common binding pocket. Considering a cutoff distance of 4 Å as the hit criterion, benchmarks carried out for weakly homologous TASSER models demonstrated a success rate of 67.3% for identifying the best of top five predicted ligand-binding sites with a ranking accuracy of 75.5%. The median sensitivity and Matthew's correlation coefficient (MCC) between predicted and observed binding residues are 0.64 and 0.59, respectively.

We also extend the application of the FINDSITE algorithm to predict metal-binding sites in weakly homologous protein models using closely as well as distantly related templates. A new approach to metal binding site prediction, FINDSITE-metal, combines structure/evolutionary information with machine learning to provide highly accurate metal binding annotations. In large-scale benchmarks against protein models constructed by TASSER, whose average Cα RMSD from the native structure is 8.9 Å, 59.5% (71.9%) of the best of top five predicted metal locations are within 4 Å (8 Å) from a bound metal in the crystal structure. In 65.6% and 83.1% of the cases, the best predicted binding site is at rank 1 and within the top 2 ranks, respectively. Furthermore, for iron, copper, zinc, calcium and magnesium ions, the binding metal can be predicted with high, typically 70-90%, accuracy. FINDSITE-metal also provides a set of confidence indexes that help assess the reliability of predictions.

The third component of SiteHunterPro carries out the DNA-binding function prediction. This component utilizes a threading-based method, DBD-Threader, for the prediction of DNA-binding domains and associated DNA-binding protein residues[3]. The method combines fold similarity and DNA-binding propensity as two functional discriminating properties. In benchmark tests on 179 DNA-binding and 3,797 non-DNA-binding proteins, using templates whose sequence identity is less than 30% to the target, DBD-Threader achieves a sensitivity/precision of 56%/86%. This performance is considerably better than the standard sequence comparison method PSI-BLAST and is comparable to methods that

require the target structure as input[9]. Additionally, DBD-Threader correctly assigns the SCOP superfamily for most predicted domains. DBD-Threader has also been validated in large-scale application on 18,631 protein sequences from the human genome.

Availability

FINDSITE, FINDSITE-metal and DBD-Threader are freely available to the academic community at http://cssb.biology.gatech.edu. Moreover they are integrated with the protein structure prediction tools, TASSER[8], TASSER-Lite[10] and METATASSER[11], into a unified web resource, Protein Structure and Function prediction Resource[12] (PSiFR), available at http://psifr.cssb.biology.gatech.edu/.

1. Brylinski,M. & Skolnick,J. (2008). A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proc Natl Acad Sci U S A* **105**, 129-34.
2. Brylinski,M. & Skolnick,J. (2008). FINDSITE-metal: integrating evolutionary information and machine learning for structure-based metal binding site prediction at the proteome level. *Proteins*, submitted.
3. Gao,M. & Skolnick,J. (2009) A Threading-based method for the prediction of DNA-binding proteins with application to the human genome, *PLoS Comp. Biol.*, **5**, e1000567.
4. Skolnick,J., Kihara,D. & Zhang,Y. (2004). Development and large scale benchmark testing of the PROSPECTOR_3 threading algorithm. *Proteins* **56**, 502-18.
5. Zhou,H. & Zhou,Y. (2004). Single-body knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. *Proteins* **55**, 1005-13.
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. Pandit,SB. & Skolnick,J. (2008). Fr-TM-align: a new protein structural alignment method based on fragment alignments and the TM-score. *BMC Bioinformatics* **9**, 531.
8. Zhang,Y., Arakaki,A.K. & Skolnick,J. (2005). TASSER: an automated method for the prediction of protein tertiary structures in CASP6. *Proteins* **61 Suppl 7**, 91-8.
9. Gao,M. & Skolnick,J. (2008). DBD-Hunter: a knowledge-based method for the prediction of DNA-protein interactions. *Nucleic Acids Res.* **36**, 3978-92.
10. Pandit,S.B., Zhang,Y. & Skolnick,J. (2006). TASSER-Lite: an automated tool for protein comparative modeling. *Biophys J*. **91**, 4180-90.
11. Zhou,H., Pandit,S.B., Lee,S.Y., Borreguero,J., Chen,H., Wroblewska,L., & Skolnick,J. (2007). Analysis of TASSER-based CASP7 protein structure prediction results. *Proteins* **69 Suppl 8**, 90-97.
12. Pandit,S.B., Brylinski,M., Zhou,H., Gao,M., Arakaki,A.K. & SkolnickJ. (2010). PSiFR: an integrated resource for prediction of protein structure and function. *Bioinformatics* **26**, 687-8.

SMEG-CCP

Prediction of Native Contacts, 3D Structures and Model Quality Using Consensus Contacts

B.H. Stehr, M. Lappe
Max-Planck-Institute for Molecular Genetics, Berlin, Germany
{stehr, lappe}@molgen.mpg.de

The SMEG-CCP approach (Sample MEan of Graphs Consensus Contact Prediction) uses contact information derived from server models to predict residue-residue contacts, 3D structures and model quality.

Methods

For each target, the server predictions marked as model 1 were converted to contact maps using the CASP contact definition of 8Å between C-beta atoms (C-alpha for glycins). The sample mean[1] of these contact maps contains for each contact the frequency of occurrence in the input ensemble. The frequencies were ranked in descending order and the top *n* were submitted as predicted contacts where *n* is the expected number of contacts in the given target. To determine *n*, we again used a consensus approach choosing *n* as the median number of contacts predicted in the server models.

From the predicted contact maps we calculated 3D models with the DISTGEOM[2] program from the TINKER[3] package. Each contact was translated into a distance restraint with 2.6Å lower and 8Å upper bound between C-beta atoms.

The quality of server models was predicted based on the agreement of their contact map with the sample mean. For each contact present in a model, the number of structures in the ensemble that share that same contact gives an estimation of the likelihood of this contact being native. These values summed over all contacts in a model give the raw quality score for the model.
Raw scores from a training set were fitted to GDT scores to derive quality estimates in terms of GDT.

Results

In many cases, the predicted contact maps are closer to native than the best input model (see e.g. Figure 1). Benchmarking on Casp8 targets has shown that an average, the consensus contact prediction is superior to any of the server methods in terms of prediction accuracy and coverage. Our method works particularly well for medium-difficulty targets where enough consensus information is available but agreement between models is not too high.

The 3D models generated from the contact predictions score lower in GDT_TS score because of inaccuracies in the atomic details resulting from the reconstruction procedure. However, the quality of the contact prediction suggests that current 3D prediction methods could be improved by incorporating consensus contact information.

Availability

A Java implementation of the method is available from the authors upon request.

Figure 1: Consensus contact prediction of T0409. The prediction (b) is derived from the ensemble (a) by the method described above. The native contact map (c) is shown for comparison.

1. Jain, B. and Obermayer K. (2008). On the sample mean of graphs. In *IJCNN 2008 Conference Proceedings. 993-1000.*
2. Hodsdon,M.E., Ponder, J.W. and Cistola,D.P. (1996). The NMR Solution Structure of Intestinal Fatty Acid-binding Protein Complexed with Palmitate: Application of a Novel Distance Geometry Algorithm. *J. Mol. Biol.* ***264***, 585-602.
3. Ponder,J.W. and Richards,F.M. (1987). An Efficient Newton-like Method for Molecular Mechanics Energy Minimization of Large Molecules, *J. Comput. Chem.* ***8***, 1016-1024.

Splicer
Splicer_QA

SPLICER: An autonomous model quality assessment method using non-linear/linear combinations of some potential energies containing statistical potential, physics-based potential and residue-residue distance potential

Yuuki Nakamura[1], Kazuhiko Kanou[1,2], Genki Terashi[1], Makoto Oosawa[1] , Hideaki Umeyama[1] and Mayuko Takeda-Shitaka[1]
[1] - School of Pharmacy, Kitasato University
2- Infectious Disease Surveillance Center, National Institute of Infectious Disease
shitakam@pharm.kitasato-u.ac.jp

SPLICER is an autonomous model quality assessment method which does not require the many server models. SPLICER combines some potential energies such as two kinds of the statistical potential, the physics-based potential and residue-residue distance potential using non-linear and linear regression methods. For the combination of the different types of scores, the non-linear regression method was employed. In our method, *spline* function was used as the no-linear regression method for the combination of the two kinds of statistical potentials and physics-based potential composed of four terms. The dataset for the regression was created using homology modeling for about six thousand target proteins whose structures were already known. By using the *spline* function with this dataset the predicted GDT_TS value for a model was obtained. The predicted GDT_TS called *sGDT_TS* has correlation with real value of the GDT_TS. Furthermore, the residue-residue distance potential (RRDP) was added to the *sGDT_TS* value for the consideration of the information of the template proteins. Then the RRDP score was normalized to the dimension of the GDT_TS value. The normalized score of the RRDP called *rGDT_TS* is combined with *sGDT_TS* value using the linear regression. This *srGDT_TS* value has stronger correlation with real value of the GDT_TS than each predicted GDT_TS value alone.

Methods

Statistical potential

SPLICER used two kinds of the statistical potentials. One is a CIRCLE[1] score which is a 3D-1D score modified based upon verify3D[2]. The CIRCLE score was calculated from 3D coordinates of a protein model including side-chain atoms, and the score estimates the stability of the protein model from the free energy point of view. The CIRCLE score was based on the following equation,

$$SCORE(AA|env) = log\left(\frac{P(AA|env)}{P(AA)}\right) \qquad (1)$$

Another is *SSscore*[1] which represents the secondary structure agreement. This score was calculated by comparison between the secondary structure of the 3D model and the secondary structure predicted from the sequence. The secondary structure prediction from the sequence was performed using PSI-PRED[3]. The measure of similarity in secondary structures is based on the following scoring function.

$$SCORE(i,j) = log\left(\frac{P(i,j|conf)}{P_{pre}(i|conf)P_m(j|conf)}\right) \qquad (2)$$

Physics-based potential

The physics-based potential comprises four terms, *Vhp, Vhb, Vcoli* and *Vrama* corresponding to potential energy for hydrophobicity, hydrogen bonds, collisions and main chain torsion angles, respectively. The physics-based potential was calculated from the 3D-coordinates of all atoms including side-chain atoms.

Residue-residue distance potential

Residue-residue distance matrix was calculated from the coordinates of all CA atoms of a model. Same distance metric were calculated from template proteins which were detected by several homology search programs for each targets. Distance matrix of the model was compared with those of template proteins, and then the score which represents the similarity of protein fold between the model M and templates was calculated as shown in Equation (3).

$$RRDP(M) = \sum\sum \frac{1}{\left(d_{M[j]} - d_{Ti[j]}\right)^2 + 1} \qquad (3)$$

Non-linear regression

Non-linear regression method was employed to combine different types of potential energies. The statistical and the physics-based potentials were combined by using *spline* function which implements a non-linear regression. Then, 446,717 homology models for many target proteins were used as a dataset for the regression. These models were constructed with FAMS[4] based on many alignments generated by several alignment programs such as PSI-BLAST[4]. The dataset includes the values of CIRCLE score, *SSscore*, *Vhp*, *Vhb*, *Vcoli*, *Vrama*, model length and real value of GDT_TS for each model. Using *spline* function with this dataset, the GDT_TS value was predicted for a protein model whose native structure is unknown. The predicted GDT_TS in this step, called *sGDT_TS*, was calculated with R program[5] using the gam (Generalized Additive Models) function as shown in Equation (4).

$$sGDT\_TS = pred\,[gam\{s(CIRCLE) + s(SSscore) + s(V_{HP}) + s(V_{HB}) + s(V_{coli}) + s(V_{rama}) + s(len_{mdl})\}]/(4*target\ length) \qquad (4)$$

Linear regression

In this step, the *sGDT_TS* was combined with RRDP score using the linear regression. Then the RRDP score was normalized to the same dimension with the GDT_TS for the purpose of the combination with the value of the *sGDT_TS*. As dataset for the linear regression, 23,584 server models in the CASP8 were used. Then the linearly combined score (*srGDT_TS*) was calculated as follows,

$$srGDT\_TS = \frac{sGDT\_TS + k * rGDT\_TS}{(k + 1)} \qquad (5)$$

Here, *rGDT_TS* is the normalized score of RRDP. *k* is a coefficient value for *rGDT_TS*, which was determined with the training dataset of the CASP8 server models. In the CASP9, we calculated the *srGDT_TS* values for the server reconstructed models which were refined by the FAMS program for the

purpose of the removal of atom collisions. Then the *srGDT_TS* value divided by 100 was submitted as a QA score for each model.

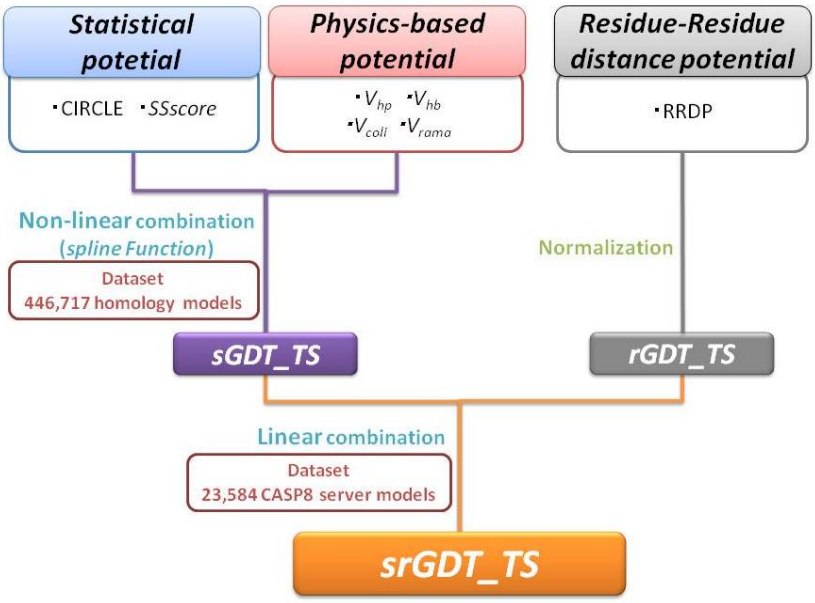


Figure 1. Schematic of SPLICER.

Results

We evaluated the performance of SPLICER using 85 experimental structures of 129 CASP9 targets became available by September 12, 2010. Table 1 shows the Pearson's *r* and the Kendall's *t* for the average and overall correlation. Also the GDT_TS loss value was shown in Table 1. Unfortunately, it was found that there are some bugs in the process of model reconstruction. 136 out of 24,073 server models (about 0.5%) were not reconstructed with propriety. The parenthetic values in the Table 1 represent the results which was fixed these bugs. Figure 2 shows the real GDT_TS value for all server models of the 93 targets plotted against the QA score. Green plus point indicates the bug models.

Average correlation	Pearson	0.847 (0.852)
	Kendall	0.601 (0.603)
Overall correlation	Pearson	0.900 (0.902)
	Kendall	0.721 (0.722)
Loss of GDT_TS		7.528

Table 1. Summary of results for Splier

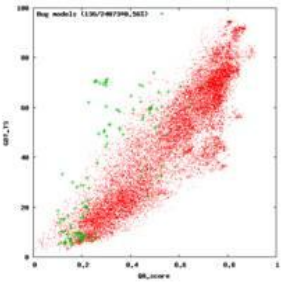


Figure 2. Real GDT_TS value plotted against QA score

1. Terashi G, Takeda-Shitaka M, Kanou K, Iwadate M, Takaya D, Hosoi A, Ohta K, Umeyama H. Proteins. 2007;69 Suppl 8:98-107

2. Eisenberg D, Lüthy R, Bowie JU. Methods Enzymol. 1997;277:396-404.
3. Jones DT. J Mol Biol. 1999 Sep 17;292(2):195-202.
4. Ogata, K. and Umeyama, H. J Mol Graph Model 2000; 18, 258-272.
5. The R project homepage : http://www.r-project.org/

Spritz3

Spritz3: protein disorder prediction using five in-house sequence predictors.

Ian Walsh[1], Alberto J. Martin[1], Gianluca Pollastri[2] and Silvio Tosatto[1]
[1] – University of Padova, [2] – University College Dublin
ian.walsh@bio.unipd.it

Non-folding flexible regions within a protein are known as disordered regions. Disordered regions are widespread within known protein structures, especially in eukaryotic organisms[1,2]. Disorder plays a key role in: human disease[3], DNA binding [4], molecular recognition[5] and functional binding sites[6].

Predictions at CASP9 were made using five methods for determining disorder. The final predictor uses a simple average of the individual system probabilities. The disorder decision threshold was determined on protein sequences which were used for learning by the individual systems. We term the final combination predictor Spritz3.

Methods

Each method was developed in-house which differs from the usual meta-servers often used at CASP and elsewhere. The five models are based on simple properties at the sequence level. Information includes: multiple sequence alignments, secondary structure, solvent accessibility, codon diversity, electrostatics, molecular volume and polarity. Each system uses a subset of this information while the final combination should see all the information. Homology is used for three of the systems, two systems use weighted homology information as input and one as a post-filter.

The datasets used for training the algorithms were all pre-CASP8 target release. This allowed us to benchmark Spritz3 on CASP8 targets where the distribution of homologues found for the targets is a realistic CASP setting (using a psi-blast search).

The sequential and structural information from the homologues are used for training Support Vector Machines. We found that three of the predictors displayed complementary probabilities (Pearson correlation is small for 3 of the predictors). Thus an ensemble average of the predictors should be partially orthogonal improving the results compared to the individual predictors[,7,8].

The method which worked best was a simple average of the probabilities with the cutting threshold of 0.15 on the training set. Regular expressions and filtering criteria were also used since they improved the performance slightly.

Results

We extensively benchmark the final predictor on CASP 8 (103 X-Ray and 19 NMR). Initial results show that a combination of the methods is improved over currently available methods and state-of-the-art performances were achieved on the CASP8 dataset. We also proved that a simple average of probabilities improves over majority and unanimous voting.

Availability

A server will be operational in the coming weeks at: http://protein.bio.unipd.it/cspritz/. At the moment the previous version of our server is available at http://protein.bio.unipd.it/spritz/.

1. Intrinsic Protein Disorder in Complete Genomes. Dunker AK, Obradovic Z, Romero P, Garner EC

and Brown CJ. Genome Informatics 2000;11:161-71.
2. Function and structure of inherently disordered proteins. A Keith Dunker, Israel Silman, Vladimir N Uversky and Joel L Sussman. Current Opinion in Structural Biology Volume 18, Issue 6, December 2008, Pages 756-764.
3. Intrinsically Disordered Proteins in Human Diseases: Introducing the D2 Concept. Vladimir N. Uversky, Christopher J. Oldfield and A. Keith Dunker. Annual Review of Biophysics, Vol. 37: 215-246 (Volume publication date June 2008) .
4. Folding transition in the DNA-binding domain of GCN4 on specific binding to DNA. Weiss MA, Ellenberger T, Wobbe CR, Lee JP, Harrison SC and Struhl K. Nature. 1990;347:575–578.
5. Calmodulin signaling: Analysis and prediction of a disorder-dependent molecular recognition. Predrag Radivojac, Slobodan Vucetic, Timothy R. O'Connor, Vladimir N. Uversky, Zoran Obradovic, A. Keith Dunker. Proteins: Structure, Function, and Bioinformatics Volume 63 Issue 2, Pages 398 - 410.
6. Prediction of Protein Binding Regions in Disordered Proteins. Bálint Mészáros, István Simon, and Zsuzsanna Dosztányi. PLoS Comput Biol. 2009 May; 5(5): e1000376.
7. Using Diversity in Preparing Ensembles of Classifiers Based on Different Feature Subsets to Minimize Generalization Error. Machine Learning: ECML (2001), Zenobi, G. and Cunningham, P.
8. Improved Disorder Prediction by Combination of Orthogonal Approaches. PLoS ONE 4(2): e4433. (2009). Schlessinger A, Punta M, Yachdav G, Kajan L, Rost B.

Sternberg

Protein structure and binding site prediction using Phyre2 and 3DLigandSite

L.A. Kelley[1*], M.N.Wass[1*] and M.J.E. Sternberg[1]
[1] – *Imperial College London*
**Authors contributed equally to this work*
l.a.kelley@imperial.ac.uk, mark.wass04@imperial.ac.uk

Methods

Human 3D structure predictions were made using a combination of structural clustering with a modified version of the Poing[1] *de novo* modeling tool as described in the Phyre2[2] CASP9 abstract. Server models were downloaded from the CASP website and clustered using our in-house maxcluster program and ranked using the 3DJury[3] protocol. High ranking models that shared significant similarity by visual inspection were then selected and used as input to the Poing modeling tool. These models provided distance constraints for the Poing simulation. In cases where multiple equally plausible yet structurally dissimilar models from different servers were produced, up to 5 runs of poing with different combinations of input models were performed.

Human binding site predictions were made using a consensus approach with our 3DLigandSite[4] server. The CASP9 server predictions were downloaded and clustered using 3DJury[3]. The top 6 models (obtained from different groups) were individually run through 3DLigandSite. The results were manually combined with those obtained using our own Phyre2 models. Individual residues were predicted to form part of the binding site based on the number of 3DLigandSite runs that had predicted them, their conservation score and on visualization of the modeled protein and the clustered ligands. Additional functional information for the targets was sought from UniProt[5], Pfam[6], Interpro[7] and ConFunc[8] to aid the manual process particularly to determine the likely ligands of the target.

1. Jefferys,B.R., Kelley,L.A. and Sternberg, M.J.E. (2010). Protein folding requires crowd control in a simulated cell. *J. Mol. Biol. 397*, 1329-1338.
2. Kelley,L.A. and Sternberg,M.J.E. (2009). Protein structure prediction on the web: a case study using the Phyre server. *Nature Protocols.* 4, 363-371.
3. Ginalski,K., Elofsson,A., Fischer,D., & Rychlewski,L. (2003) 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics* 19, 1015-1018.
4. Wass,M.N., Kelley,L.A., & Sternberg,M.J. (2010) 3DLigandSite: predicting ligand-binding sites using similar structures. *Nucl. Acids Res.* 38, W469-473.
5. The UniProt Consortium (2009) The Universal Protein Resource (UniProt) 2009. *Nucl. Acids Res.* 37, D169-174.
6. Finn,R.D., *et al.* (2010) The Pfam protein families database. *Nucleic Acids Research* 38, D211-D222.
7. Hunter,S., *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Research* 37, D211-D215.
8. Wass,M.N. & Sternberg,M.J. (2008) ConFunc--functional annotation in the twilight zone. *Bioinformatics* 24,798-806.

StruPPi

Homology Modeling of Protein Structure using Fragment/Profile based search method in CASP9

M.Parthiban[1], S. Susdalzew[1], S. Belsare[1], H. Stehr[1], M. Winklemann[1], C. L. Sargent[2], E. Michalsky[2], M. Schneider[2], R. Preissner[2] and M. Lappe[1]

[1] - *Max-Plank Institute for Molecular Genetics, Otto-Warburg Laboratories, Ihnestraße. 63-73, 14195, Berlin, Germany*

[2] - *Structural Bioinformatics Group, Charite Campus Buch, Lindenberger Weg 80, 13125, Berlin, Germany*

{parthi, lappe}@molgen.mpg.de

We employed two different approaches for tertiary structure prediction.

Approach 1. For targets with clearly identifiable templates.
A conventional approach to homology-based modeling was applied.

1. GenThreader[1] server (Profile-based secondary structure prediction) was used to identify potential 3D structural homologues.
2. The alignment was manually improved based on secondary structure predictions using Jalview.
3. 20 models were generated using MODELLER9V7[2], where the models were built based on spatial constraints.
4. These structures were ranked based on the DOPE[3] scoring function and then subjected to further refinement.

Approach 2. Assembly of weak distance constraints.

1. In contrast, when a meaningful template could not be identified for the target sequence, we applied a remote homologue search method by using PSI-Blast to identify distantly related homologue structures.
2. We also used consensus-based contact information of multiple templates from server models by CMView package, which uses distance geometry from the TINKER[4] package.
3. The resulting structure from the latter approach was then subjected to further refinement.

Refinement

SuperLooper[5] was used to improve regions of the models with predicted inaccuracies. It is based on a comprehensive database of protein segments from the PDB. Loop regions were then examined for the most suitable conformation. The resulting model was subjected to energy minimization by Steepest Descent/Conjugate Gradient algorithms using Accelrys Discovery Studio and thus the final structure was prepared.

Evaluation of Target Structures

Finally structures were ranked using the DOPE scoring function and secondary structure positions were improved using Ramachandran plot viewer.

Alternative conformations were scored using the DOPE score and then assigned accordingly as models TS1-TS5 for submission.

1. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. Bioinformatics. 2000; 16 (4): 404-5.
2. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen MY, Pieper U, Sali A. Comparative protein structure modeling using MODELLER. Curr Protoc Protein Sci. 2007 Nov; Chapter 2:Unit 2.9.
3. Shen MY, Sali A. Statistical potential for assessment and prediction of protein structures. Protein Sci. 2006 15(11): 2507-24.
4. Ponder,J.W. and Richards,F.M. (1987). An efficient newton-like method for molecular mechanics energy minimization of large molecules, J. Comput. Chem. 8, 1016-1024.
5. Hildebrand PW, Goede A, Bauer RA, Gruening B, Ismer J, Michalsky E, Preissner R. SuperLooper--a prediction server for the modeling of loops in globular and membrane proteins. Nucleic Acids Res. 2009 1;37

SUN_at_tsinghua

All-Atom CSAW: An Ab Initio Protein Folding Method

Weitao Sun[1]
[1] – *Zhou Pei-Yuan Center for Applied Mathematics, Tsinghua University, Beijing, China, 100084*
sunwt@tsinghua.edu.cn

All-Atom Conditioned Self-Avoiding Walk (AA-CSAW) is an *ab initio* protein folding simulation model based on Monte-Carlo (MC) method(Huang, 2007; Huang, 2008; Sun, 2007). The polypeptide chain is simulated as effectively rigid cranks $-C_a-CO-NH-$ units lined by covalent bonds. Bond lengths and bond angles are set as fixed optimal values. The structure of polypeptide is fully described by backbone dihedral angles ϕ,ψ and the sidechain dihedral angles χ. The number of χ depends on the type of amino acid. A trial structure is randomly generated by pivoting the polypeptide chain and sidechains. In the pivot algorithm, the backbone dihedral angles ϕ,ψ for each residue are chosen in Ramachandran plot according to a probability distribution derived from 3-residue fragment set. The effective energy of protein structure is constructed by considering hydrophobic effect, desolvation effect and hydrogen bonding interaction. An appropriate three dimensional structure is accepted with a probability according to Metropolis scheme(Metropolis, 1987). In order to evaluate the accepted structures in MC simulations, the ratio of secondary structure content to radius of gyration is introduced.

Methods

Backbone dihedral angle distribution

By selecting special dihedral angles ϕ_i,ψ_i for residue i, the polypeptide chain will change to a different 3D conformation. In general, ϕ_i,ψ_i can be any values in Ramachandran plot except those prohibited by steric effect. However, observations of Protein Data Bank(Berman, et al., 2000) data show that the distribution of ϕ_i,ψ_i in Ramachandran plot is far from uniform. It seems that the dihedral angle values of residue i have obvious relations with the amino acid types of residue $i-1$ and $i+1$. We constructed dihedral angle distribution models for all 20 amino acids based on a high resolution 3-residue fragment database. This prior information substantially improve the accuracy and convergence of AA-CSAW method.

Secondary structure definition

Each residue of protein can be in helix, strand, turn or coil structure. The secondary structure property (SSP) of a residue is important for monitoring the folding stage. The SSP is usually determined by hydrogen bonding interactions. In AA-CSAW, we use the algorithms described in Stride method(Frishman and Argos, 1995) .

Effective structure energy

The effective structure energy is composed of three parts: hydrophobic effect, hydrogen bonding and desolvation energy.

Hydrophobic effect

In AA-CSAW, the hydrophobicity of each residue depends on the corresponding amino acid type. The hydrophobic energy is estimated based on two factors: the solvent accessible surface area (SASA) and residue types. For residue i, if it has more neighbors, it is buried in protein and has less SASA. In addition, if the surrounding residues are all hydrophobic residues, the hydrophobic energy of residue i is high.

The 'dewetted' phenomenon near the surface between large nonpolar groups and water is considered in AA-CSAW. In conventional continuum water solvent models, hydrophobic effect is always overestimated for the reason that water molecules are more dilute near large nonpolar groups. We introduce a scheme to decrease the hydrophobic energy when the aggregation of hydrophobic residue grows to large size. This method provide more chances to open the hydrophobic core, which is essential for misfolded intermediate structures.

(a) Hydrogen bonding (HB)

Each residue carries both HB donor and HB acceptor. We scan NH, CO groups in every residue and check if these groups between residue i and j ($j \neq i \pm 1$) satisfy the HB conditions. In AA-CSAW, the DSSP (Kabsch and Sander, 1983) method is used as HB criterion. The total number of hydrogen bonds is a measurement of HB energy. Since the stability of hydrogen bond may depend on it location, a optimal HB strength parameter is used as a weight. If the hydrogen bond is buried in protein interior, the weight value is high. Otherwise, the peptide-peptide hydrogen bond is exposed to water and can be easily destroyed. Thus the weight value is low.

(b) Desolvation energy

Hydrophobic effect leads to a fast collapse of polypeptide chain. Hydrogen bonding interactions cause the emergence of secondary structures. However, a collapsed chain with hydrophobic core but without hydrogen bond is usually in high free energy state. In order to prevent the formation of tight hydrophobic core without hydrogen bonding, we introduce a penalty to buried NH, CO groups that can't form hydrogen bonds for some reasons.

Structure evaluation parameter

The AA-CSAW is now a parallel code and can produce many candidate structures. We find that the ratio of secondary structure content to radius of gyration is a pretty good indicator for evaluating a structure. This value usually depends on the length of a protein. For the same protein, the higher this ratio, the better the predicted structure.

Results

All results, intermediate data files, and performance analysis documents will soon be available on the web at http://zcam.tsinghua.edu.cn/~sunwt/aacsaw.htm.

Availability

The AA-CSAW version 1.0.0 is written in C++ and have been compiled and tested on both WindowsXP and LINUX systems. The software is to be downloaded at http://zcam.tsinghua.edu.cn/~sunwt/aacsaw.htm soon, as well as the manuals and FAQ.

1. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank, *Nucleic Acids Res*, **28**, 235-242.
2. Frishman, D. and Argos, P. (1995) Knowledge-based protein secondary structure assignment, *Proteins*, **23**, 566-579.
3. Huang, K. (2007) CONDITIONED SELF-AVOIDING WALK (CSAW): STOCHASTIC APPROACH TO PROTEIN FOLDING, *Biophysical Reviews and Letters* **2**, 139-154.
4. Huang, K. (2008) PROTEIN FOLDING AS A PHYSICAL STOCHASTIC PROCESS, *Biophysical Reviews and Letters* **3**, 1-18.
5. Kabsch, W. and Sander, C. (1983) Dictionary of Protein Secondary Structure - Pattern-Recognition of Hydrogen-Bonded and Geometrical Features, *Biopolymers*, **22**, 2577-2637.
6. Metropolis, N. (1987) The Beginning of Monte Carlo Method, *Los Alamos Science*, **15**, 125–130.
7. Sun, W. (2007) Protein folding simulation by all-atom CSAW method, *IEEE International Conference on Bioinformatics and Biomedicine*, **2**, 45 - 52.

SVMSEQ

SVMSEQ for ab initio protein residue contact prediction

Sitao Wu[1] and Yang Zhang[2]
[1]Center for Research in Biological Systems, University of California, San Diego, 9500 Gilman Dr., La Jolla, CA 92093, [2]Center for Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw Ave, Ann Arbor, MI 48109
zhng@umich.edu

We developed SVMSEQ, a machine-learning-based method for *ab initio* contact predictions, trained on a variety of sequence-derived features[1]. These include (1) *Local window features:* position-specific scoring matrices, secondary structure predictions, solvent accessibility predictions; (2) *In-between segment features:* the contact order, the compositional percentage of three different secondary structure elements and two different burial states for the in-between residues, state distributions of the in-between residues, and the local features of five selected in-between residues. In summary, for short/medium/long ranges (corresponding to the sequence separation in 6-11, 12-23 and ≥24 residues, respectively), there are 781/787/918 input features, which are used for SVM[2] to classify the contact and non-contact pairs. The top contact pairs ranked by confidence scores are submitted as the final predictions of SVMSEQ. In a recent study, it was found that multiple SVMSEQ-based contact predictions can provide significant improvement on the 3D structure assembly of non-homologous proteins[3].

In CASP8, we used a linear combination of template-based (LOMETS[4]) and *ab initio* contact prediction (SVMSEQ[1]) to generate contact predictions, which worked well for template-based targets but performed poorly for the free modeling (FM) targets. In CASP9, we used SVMSEQ to generate contact prediction for all targets.

The on-line server and the SVMSEQ package are freely available at http://zhanglab.ccmb.med.umich.edu/SVMSEQ.

1. Wu, S. & Zhang, Y. (2008) A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. *Bioinformatics (Oxford, England)*.**24**, 924-931.
2. Joachims, T. (2002) Dissertation: Learning to Classify Text Using Support Vector Machines. Software available at http://svmlight.joachims.org/.
3. Wu, S., Szilagyi, A. & Zhang, Y. (2010) Improving protein structure prediction using multiple sequence-based contact predictions. Submitted.
4. Wu, S. & Zhang, Y. (2007) LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic acids research*.**35**, 3375-3382.

SWA_TEST

Testing a StepWise 'Ansatz' for High Resolution Macromolecule Modeling

Kyle Beauchamp[1], Parin Sripakdeevong[1], and Rhiju Das[1,2,3]
[1]Biophysics Program, [2]Biochemistry Department, and [3]Physics Department, Stanford University
rhiju@stanford.edu

High-resolution structure modeling of protein, RNAs, and other macromolecular systems is difficult. A major bottleneck is the intractability of sampling these problems' many number of degrees of freedom. To ameliorate this conformational sampling issue, Rosetta approaches typically use a low-resolution sampling stage, fragments from experimental structures, or a Monte-Carlo-like search -- frequently all three[1]. Nevertheless, these approaches achieve near-atomic-resolution blind predictions only in a small number of favorable cases[2]. We have been exploring an alternative strategy for generating high-resolution Rosetta models that avoids low-resolution sampling, fragments, and stochastic search methods.

Methods

We describe a step-wise "ansatz" (SWA) that builds models in small steps, enumerating several million conformations for each residue, and covering all possible build-up paths through a dynamic-programming-like strategy. For a number of modeling problems, this deterministic method produces well-packed, well-hydrogen-bonded conformations. After implementing the method for noncanonical RNA loops, RNA tertiary contacts, and mini-proteins, we discovered this line of inquiry follows step-by-step buildup procedures explored by Levinthal[3] and, later, Scheraga[4]; and dynamic-programming aspects echo recent work by Dill and Joshi[5]. This prior work has stayed outside the mainstream of structure modeling, perhaps due to computational cost. Nevertheless, step-wise, enumerative modeling appears powerful when brought together with modern high-performance computing and physical reasonable all-atom energy functions.

Results

We present results on 12- to 20- residue non-canonical RNA motifs as well as highly irregular protein loops that have been intractable for prior fragment assembly[6], analytic loop closure[7], or hierarchical[8] approaches. In all cases, step-wise assembly either reaches atomic accuracy, exposes flaws in the Rosetta high-resolution energy function, or requires backbone flexibility outside the modeled loop. Blind tests on an RNA tetraloop-receptor motif[9] are underway, as well as extension of the method to more complex systems such as RNA aptamers and sprotein knottins. CASP9 presented numerous loop modeling tests (in comparative modeling targets) as well as at least one small-protein (~60 residue) target within the size range reachable with available computational power; we look forward to their evaluation by the CASP9 expert assessors.

Availability

The SWA method is unpublished work; it has been implemented in the Rosetta codebase and is available upon request from the authors.

1. Das, R. & Baker, D. (2008). Macromolecular modeling with rosetta. *Annu Rev Biochem* **77**, 363-82.
2. Raman, S., Vernon, R., Thompson, J., Tyka, M., Sadreyev, R., Pei, J., Kim, D., Kellogg, E., DiMaio, F., Lange, O., Kinch, L., Sheffler, W., Kim, B. H., Das, R., Grishin, N. V. & Baker, D. (2009). Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins* **77 Suppl 9**, 89-99.
3. Levinthal, C. (1968). Are there pathways for protein folding. *Journal de Chimie Physique et de Physico-Chemie Biologique* **65**, 44-45.
4. Vasquez, M. & Scheraga, H. A. (1985). Use of buildup and energy-minimization procedures to compute low-energy structures of the backbone of enkephalin. *Biopolymers* **24**, 1437-47.
5. Hockenmaier, J., Joshi, A. K. & Dill, K. A. (2007). Routes are trees: the parsing perspective on protein folding. *Proteins* **66**, 1-15.
6. Wang, C., Bradley, P. & Baker, D. (2007). Protein-protein docking with backbone flexibility. *J Mol Biol* **373**, 503-19.
7. Mandell, D. J., Coutsias, E. A. & Kortemme, T. (2009). Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nat Methods* **6**, 551-2.
8. Jacobson, M. P., Pincus, D. L., Rapp, C. S., Day, T. J., Honig, B., Shaw, D. E. & Friesner, R. A. (2004). A hierarchical approach to all-atom protein loop prediction. *Proteins* **55**, 351-67.
9. Costa, M. & Michel, F. (1997). Rules for RNA recognition of GNRA tetraloops deduced by in vitro selection: comparison with in vivo evolution. *EMBO J* **16**, 3289-302.

TASSER

TASSER for protein structure prediction in CASP9

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

TASSER human expert group has submitted predictions for protein structure and refinement in CASP9. We have tested some new methods in this CASP compared to the last one. Those new methods include FTCOM[1] for selecting models from CASP servers, TASSER_WT[2] for generating more accurate contacts for Medium/Hard targets. We also implemented a new way of building all atom models from the C_α-only models from TASSER[3] simulations.

Method

Targets were classified into Easy, Medium and Hard categories if the Z-score of the first SP[3] threading template is >6.0, 4.5 <= Z-score <=6.0 and < 4.5 respectively. For structure prediction, we selected models from all the CASP servers with TASSER-QA[4] for Easy targets as in CASP8 and the newly developed FTCOM[1] method for Medium/Hard targets. FTCOM is a method for ranking models by comparing models to both top ranked templates and 9-residue sliding window fragments. Tertiary restraints and contacts are then derived from those selected models. For Medium/Hard targets, additional chunk models from chunk-TASSER[5] are also included in deriving restraints and contacts. Furthermore, contacts generated by TASSER_WT[2] are combined with those derived from the models for Medium/Hard targets. TASSER-WT provides confidence weighted contacts from two variants of PROSPECTOR threading and a new fragment based threading approach that stitches together predicted local fragments to provide for better models of Medium/Hard targets. The distance restraints and contacts are then fed into TASSER to refine the selected models. We performed a single long simulation of chunk-TASSER followed by SPICKER[6] clustering for each target. Top five cluster centroid models are selected. The models only contain Cαs and usually contain Cα atom clashes and have bad geometry. We fix these problems by rebuilding the full backbone with ideal bond lengths and bond angles starting from the TASSER model that is closest to the cluster centroid. We then relax the built models using the Cα-only model as a constraint and energy functions that contain all TASSER's energy terms and an H-bond score given by the number of hydrogen bonds. Side-chains are built on those relaxed ideal geometry models with an in-house template-based approach. For each target, the top five template alignments are used for side-chain building. Starting from the top template model, if the aligned template residue is identical to the target, the side-chain rotamers of the template are copied to the target. For those residues in the target without an identical aligned residue in any of the five templates, we build the side-chains by optimizing the DFIRE[7] energy function with a simple sampling procedure by changing side-chain conformations sequentially along the chain.

We have used two methods for target refinement. One is loop modeling by generating a large number of alternative loop conformations based on the information provided by CASP organizers and using the DFIRE energy function to select final models. The other is an all-atom refinement protocol by sampling only the loop or tail regions provided by organizer using TASSER energy and an H-bond score implemented in all atom representation. Final models are selected with TASSER-QA. Our submission contains both kinds of refined models.

Result

The predicted structures by TASSER human expert have better quality than our server predictions by about 6% of TM-score for the first models of the released 35 human targets (by Sept. 7, 2010). This is mainly due to the better pool of structures from the CASP servers. Compared to the last CASP, our models have better geometry and better H-bond score and side-chain accuracy according to our partial assessment. One particular successful example of our refinement protocol is TR622. The starting RMSD from native is 7.5 Å, our 4th submitted model has a 3.9 Å. However, one average, the submitted first models are still worse than the starting models.

Availability

TASSER programs as well as their related services are available through our webpage at http://cssb.biology.gatech.edu/

1. Zhou,H and Skolnick, J. (2010) Improving threading algorithms for remote homology modeling by combining fragment and template comparisons. Proteins. **78**, 2041-8.
2. Lee,SY and Skolnick,J. Submitted. TASSER_WT: A protein structure prediction algorithm with accurate predicted contact restraints for difficult protein targets. Biophysical Journal.
3. Zhang, Y. and J. Skolnick (2004) Automated structure prediction of weakly homologous proteins on genomic scale. Proc. Natl. Acad. Sci. (USA) **101**,7594--7599.
4. Zhou,H. and Skolnick,J.(2007) Protein model quality assessment prediction by combining fragment comparisons and a consensus Cα contact potential. Proteins **71**,1211--1218.
5. Zhou, H and Skolnick, J. (2007) Ab initio protein structure prediction using chunk-TASSER.. Biophysical Journal. **9**3,1510-8.
6. Zhang, Y. and Skolnick,J. (2004) SPICKER: a clustering approach to identify near-native protein fold. J. Comput Chem **25**, 865--871.
7. Zhou, H. and Zhou,YQ. (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. PROTEIN SCIENCE. 12, 2121-2

Taylor

CASP9 predictions in the Taylor lab: Manual and Fully Automated Hybrid Modelling with Templates and Ideal Forms.

M.I. Sadowski,1 J.W. Saldanha1, K. Maksimiak1, J.T. Macdonald1,2,3, and W.R. Taylor1

1 - Division of Mathematical Biology, MRC National Institute for Medical Research, The Ridgeway, Mill Hill, London, NW7 1AA UK.

2 - Centre for Synthetic Biology and Innovation, Imperial College London, London, SW7 2AZ, UK

3 - Division of Molecular Biosciences, Imperial College London, London, SW7 2AZ, UK

msadows@nimr.mrc.ac.uk

Taylor group predictions for CASP9 used a mixed template-based and *de novo* strategy. The results of the two fully-automated servers, PROTAGORAS and PLATO were combined with a clustering-based assessment of model quality to determine high-difficulty targets, for which manual predictions were made.

Methods

Targets for which the template-based server PROTAGORAS did not identify a high-coverage template were identified as potentially difficult. For these targets an assessment of target difficulty was made based on all server models: all-against all comparisons were made with TM-align [1] to generate a mean score for each model. A difficult target was assigned as one for which the best full-coverage model had a TM-score < 0.5.

For difficult targets manual template identification was attempted based on low-confidence PSIBLAST [2] and HHPred [3] results with reference to secondary structure predictions using PSIPRED [4] and functional reasoning. Where no template could be identified the results of the PLATO server were combined with manual *ab initio* predictions using the ideal forms and the top five models were chosen to represent low-energy models as ranked by the Dfire protential [5].

1. Y. Zhang, J. Skolnick (2005) *TM-align: A protein structure alignment algorithm based on TM-score* , *Nucleic Acids Research* **33**: 2302-2309
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. Soding J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**, 951-960
4. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* ***292***, 195-202.
5. Yang T., Zhou Y. (2008) Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions *Protein Science* **72**, 1212-1219

TMD3D

Protein Hub; Automatic Protein Structure Prediction & Optimization System

Taeho Jo, Hiroshi Tanaka
1- Dept. of Bioinformatics, Medical Research Institute, Tokyo Medical and Dental University

jotaeho@bioinfo.tmd.ac.jp; tanaka@cim.tmd.ac.jp

Protein-hub is a fully automated software package that implements a homology modeling and optimization based on a measure of hydrogen bonds potential. With Protein-hub, we participated in CASP9 as a human expert team 'TMD3D' in Tertiary Structure (TS) and Quality Assessment (QA) categories.

Methods

Top templates were investigated and selected manually using PSI-BLAST[1] and HHsearch[2]. With these templates, Protein-hub generated the target secondary structures by PSIPRED[3] and calculated secondary structures of the selected templates by DSSP[4]. The initial 3D models were built automatically from these results by MODELLER 9v7[5].
These predicted results were refined by optimizing algorithm within Protein-hub. In this stage, the potential energies of cooperative hydrogen bonds were calculated and compared. Three separated packages (energy.hydrogenBond; energy.hydrogenBodsPairs; energy.Hydrogen BondsAngle) in MESHI[6] are implemented. Monte-Carlo-Minimization combined with the Linear-BFGS minimization method was used for the enhancement of beta-sheet assembly[7].

Availability

The Protein Hub server can be access from the following URL: http://www.proteinhub.net .

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. Soding, J. (2005). Protein homology detection by HMM-HMM comparison. Bioinformatics, 21, 951-960.
3. Jones,D.T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292, 195-202.
4. Cabsch,W. & Sander,C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers, 22, 2577-2637.
5. Sali,A. & Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234, 779-815.
6. Kalisman, N., Levi, A., Maximova, T., Reshef, D., Zafriri-Lynn, S., Gleyzer, Y. & Keasar, C. (2005). MESHI: a new library of Java classes for molecular modeling. Bioinformatics, 21, 3931-3932.
7. Levy-Moonshine, A., Amir, E.D., Keaser, C. (2009). Enhancement of beta-sheet assembly by cooperative hydrogen bonds potential. Bioinformatics, 25, 2639-2645.

United3D

United3D: Combination of consensus QA methods

Mayuko Takeda-Shitaka[1], Makoto Oosawa[1], Kazuhiko Kanou[1,2], Yuuki Nakamura[1] and Genki Terashi[1]
[1] - School of Pharmacy, Kitasato University
2- Infectious Disease Surveillance Center, National Institute of Infectious Disease
shitakam@pharm.kitasato-u.ac.jp

United3D is a simple Model Quality Assessment Program using two type of consensus method. The first is based on optimized clustering with similarity of each model (similar to median MaxSub, median GDT_TS and 3D-Jury[1]). The second is based on the results of our residue-residue contact prediction considering the conservation of all contacts among server models. We also tried Tertiary Structure prediction and Disorder Prediction. These our predictions were simply based on our QA method. All predictions of United3D were obtained from fully automated procedures.

Methods

Quality Assessment (QA)

As described above, United3D carried out clustering and residue-residue contact prediction. The QA score is described as:

$$QAscore = S_{clustering} + w(STD, S_{clustering}) \cdot S_{contact} \quad (1)$$

where $S_{clustering}$ is the consensus score from clustering, $S_{contact}$ is the log-odds score based upon our contact prediction, *STD* is a standard deviation of server models obtained from clustering, and *w* is a weighting function. All of parameters used in United3D were optimized from CASP8 data with Kendall's *tau* and Pearson's *r*.

Tertiary Structure prediction (TS)

According to our QA results, top 20 models were selected among server models for constructing ~50 new candidate models. The candidate models were generated by swapping exposed regions (excluding core of model). The side chains of the generated models were optimized by SCWRL[2]. Finally, the generated models were re-ranked by our QA method.

Disorder Prediction (DP)

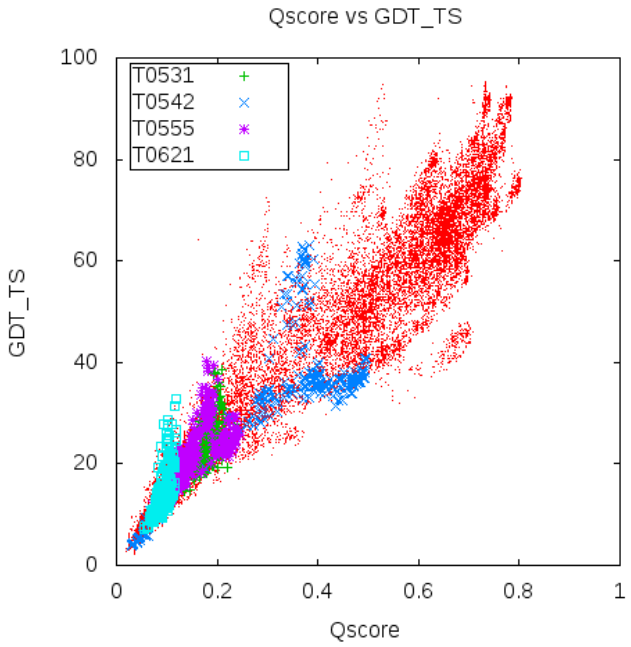
Each residue was represented by a feature vector. The vector contains PSIPRED[3] output, Unied3D local quality score and DISOPRED[4] output. The feature vectors were fed into Support Vector Machines (SVM) to predict disorder regions. The output of SVMs was filtered and normalized.

Results

According to our preliminary assessment based on the released 85 experimental structures, United3D shows slightly better results than basic clustering method (Median MaxSub) in correlation of both average and overall. In addition to the results of QA, United3D can also select better models as top 1, although United3D was not optimized with top GDT_TS. For 4 targets (T0531, T0542, T0555 and T0621), United3D did not perform well ($r < 0.7$). This is likely because the clustering method could not work well in hard targets and the case when the largest cluster does not contain good models.

TABLE I. Preliminary analysis on 85 targets of CASP9

		United3D	Median MaxSub
Average	Pearson's *r*	0.928	0.921
	Kendall's *tau*	0.682	0.667
Overall	Pearson's *r*	0.938	0.927
	Kendall's *tau*	0.793	0.777
Average GDT_TS of top1 ranked model		58.91	58.76
Average loss of GDT_TS		5.563	5.714



Availability

More detail information and results will be available on the our web site at http://pharm.kitasato-u.ac.jp/bmd/

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. Canutescu,A.A., Shelenkov,A.A. & Dunbrack R.L. (2003). A graph theory algorithm for protein side-chain prediction. Protein Sci. 12, 2001-2014.
3. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol.* **292**, 195-202.
4. Jones, DT and Ward, JJ. (2003) Prediction of disordered regions in proteins from position specific score matrices. *Proteins*, **53**, 573–578.

WAC_Labs

Fold Recognition of Highly Divergent Protein Sequences using Fold-Specific PSSM Libraries

Yoojin Hong[1,2], Sree V Chintapalli[2], Gaurav Bhardwaj[2], Damian B. van Rossum[2*] and Randen L. Patterson[2*]

[1] – *Department of Computer Science and Engineering, The Pennsylvania State University*

[2] – *Center for Computational Proteomics, The Pennsylvania State University*

[*] *To whom correspondence should be addressed: rlp25@psu.edu; dbv10@psu.edu*

Accurately recognizing structurally homologous folds for divergent protein sequences is the first step for modeling 3-D structures of proteins by comparative modeling. We recently developed a new fold recognition method[7] using Position-Specific Scoring Matrices (PSSM) libraries, whereby we generated a library for each fold contained in the SCOP database(1,086 folds, SCOP version 1.65)[3]. Given a query sequence, our method calculates fold-specific scores for these SCOP folds based on the alignments between the query sequence and the PSSM libraries obtained from rps-BLAST[1,2]. We observe that when protein sequences are represented as vectors of fold-specific scores, distant relationships can be inferred based on the correlation between the vectors. In this study, we applied our method to identify the best template structures for the queries provided in the CASP9 competition. In doing so, we determined that (i) the SCOP fold database is incomplete, (ii) that manipulating the e-value of rps-BLAST provides a robust method for accurately detecting homologous folds in the "twilight-zone" of sequence similarity, and (iii) that the Protein Data Bank provides a rich source of new fold-specific information for our PSSM libraries.

METHODS

To determine which fold group(s) the sequences were homologous to, target sequences were first screened at e-value 0.01 in rps-BLAST using a fold-specific PSSM library built by PSI-BLAST[1] with fold-specific domains in SCOP 1.65 database and their expanded sequences. A fold-specific score was calculated for each positive fold. For sequences not obtaining fold-specific scores over 1, the process was repeated at e-value $1 \times 1e10$. From these results, sequences were then hierarchically clustered into the appropriate fold-specific PSSM library. The SCOP fold structures having the highest Pearson's correlation values were then aligned to the target sequence using MUSCLE, followed by ClustalX[4] realignment in some cases (Jalview 2.5). Modeller[5] was then used to generate threaded structures. In cases where multiple structures of sufficient similarity could be identified, the multiple template option was used. Spare-parts for structures were obtained using a PSI-BLAST search of the PDB[6] (e-value=$1 \times 1e-6$).

RESULTS

For this experiment, we used the 1,086 folds defined in SCOP database as reference sequences. Given this resource, we determined that (a) we were able to successfully model a number of the targets in the human-expert CASP9 targets, and (b) that there were a number of CASP9 targets for which homologous folds were not present in the SCOP database. For example, we observed that our models for the targets T0520 and T0523 accord well to the crystalline structures, with the carbon-backbones of our models deviating less than 1.5 angstroms from observed. Through further inspection of the results, we determined that the PDB contains many fold-groups that are not present in SCOP, and when we created PSSM libraries for them, they were highly predictive. We also determined that the fold-specific score collected from a single e-value are prone to error, while collecting scores from multiple e-values can be highly accurate. While our method is not yet perfect, constructing PSSM libraries for all unique clusters

of similar protein structures in the PDB is a promising avenue to pursue towards solving the inverse-fold problem.

AVAILABILITY

The 1,086 fold-specific PSSM libraries that are used in our experiment and all codes are available upon request.

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. Schaffer, A.A., Wolf, Y., Ponting, C.P., Koonin, E.V., Aravind, L., Altschul, S.F. (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics* **15**:1000-1011.
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. Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G. (1997) The CLUSTALX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25**:4876-4882.
5. Sali, A., Blundell, T.L. (1993) Comparative protein modeling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**:779-815.
6. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., et al. (2000) The Protein Data Bank. *Nucleic Acids Res.* **28**:235-242.
7. Hong, Y., Ko, K.D., Bhardwaj, G., Zhang, Z., van Rossum, D.B., Patterson, R.L. (2010) Towards solving the inverse protein folding problem. *Physics archives* arXiv:1008.4938:q-bio.QM.

Wolfson-serv

Protein Structure Prediction using a Docking-Based Hierarchical Folding Scheme

I.Kifer[1], R.Nussinov[2] and H.J.Wolfson[1]
[1] - School of Computer Science, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, 69978 Israel
[2] - Department of Human Molecular Genetics and Biochemistry, Sackler Institute of Molecular Medicine, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv 69978, Israel
wolfson@tau.ac.il

Our server attempts to address the protein structure prediction problem using a docking based approach, based on the premise that proteins fold hierarchically[1]. Following this notion, we first attempt to assign the target sequence with structural fragments and then deduce their relative orientation towards each other. The approach was first introduced for constructing templates for High-Accuracy targets[2], and was now enhanced to address TBM targets as well[3].

Methods

Our method, FOBIA, is divided into a preprocessing stage and two online stages – an assignment stage and an assembly stage. In the preprocessing stage a non-redundant version of the PDB was cut into building blocks – sequentially consecutive structural fragments that are stable in solution[4]. These are structurally clustered, and an HMM profile is built for every cluster using the HHPred suite[5].

In the assignment stage an HMM is built for the target sequence and searched against the building block HMM dataset. Building block clusters are scored using an SVM-based machine learning approach. A graph theoretic approach is used to compute a set of building block paths that cover the target sequence as best as possible. Paths are clustered by structural similarity.

In the assembly stage we use the top two scoring paths from the previous stage. We try to deduce building block orientation by superimposing the path building blocks on a template that is most structurally related to the target sequence. We first identify such templates using HHpred and then re-rank them using FiberDock[6-7], a docking refinement algorithm, by optimizing side-chain placement and scoring the relative building block orientation energetically. The top ranking templates are chosen by this score. These templates (denote *assisting template*s) are used to orient the building blocks towards each other, as well as to fill unassigned regions of the target sequence

Given a path and an assisting template T, a model is generated as follows: we replace the relevant parts of T's structure with the path building blocks. We generate an alignment by replacing parts of the alignment between the target and T with the alignment of the target to the relevant building blocks. This hybrid template and alignment are the input to MODELLER.

Results

This approach is being tested blindly in CASP for the first time. Throughout most of the experiment the method was still being developed and fine tuned.

We have shown that our template-ranking procedure shows promising results in comparison to the HHpred method on the CASP8 banchmark[3]. We expect to learn much from this experiment towards improving our method.

Availability

Our server is available at http://bioinfo3d.cs.tau.ac.il/FOBIA/ . Results are sent to the user by email.

1. Lesk A.M. and Rose G.D. (1981) Folding unit in globular proteins *Proc. Natl. Acad. Sci.* **78**:4304-4308
2. Kifer, I., Nussinov, R. and Wolfson. H.J. (2008) Constructing templates for protein structure prediction by simulation of protein folding pathways. *Proteins*, **73(2)**:380-394
3. Kifer I., Nussinov R. and Wolfson H.J.. Protein structure prediction using a docking-based hierarchical folding scheme. *In Submission*
4. C.J. Tsai, J. Maizel, and R. Nussinov (2000). Anatomy of protein structure: Visualizing how a 1d protein chain folds into a 3d shape. *Proc. Natl. Acad. Sci.* **97**, 12038-12043
5. Söding J.(2005) Protein homology detection by hmm-hmm comparison. *Bioinformatics* **21**:951-960
6. Andrusier N., Nussinov R. and Wolfson H. J. (2007) Firedock: Fast interaction refinementin molecular docking. *Proteins*, **69**(1):139-159
7. Mashiach E., Nussinov R. and Wolfson H.J. (2009) Fiberdock: Flexible induced-fit backbone refinement in molecular docking. *Proteins* **78**(6):1503-1519

Wolynes

Structure Predictions with the Associative Memory Hamiltonian

Ryan MB Hoffman, Vanessa Oklejas, Patricio O Craig, Nicholas P Schafer, Benjamin Madej, and Peter G Wolynes
Center for Theoretical Biological Physics, University of California, San Diego
pwolynes@ucsd.edu

Our group focuses on correctly predicting overall folds, as our structure prediction approach amounts to simulated annealing under a very coarse grained (but physically meaningful) hamiltonian.[1,2] The potential here was the same as used for previous CASPs; the Associative Memory hamiltonian with Water-mediated contacts (AMW). [6] Our performance this year, if substantially different from previous years, probably stems from the use of a slightly different model selection procedure, the use of different experimenters, and luck.

AMW is one of the latest hamiltonians produced by our group over the last 20 years. [1,2,3,4,5,6] Most of these can be used for structure prediction through the global alignment of the query sequence on a library of memory proteins.[4] This gives a residue index-specific energy function, which specifies pairwise contact energies. In the case of AMW, both sequence-independent and sequence-dependent interactions are parametrized. Water mediated contacts are instantiated in a residue-specific way; for example, two residues having a like charge are repulse in a high-density (desolvated) environment but favorable in a low-density (solvated) environment. As well, to improve the accuracy of our results, we also use the Jpred 3 algorithm to determine appropriate backbone dihedral restraints. [7]

Once the energy function is defined for a particular query protein, we subject a random (and extended) initial conformation to molecular dynamics, under continuous cooling. The length of the cooling protocol varied with the system and the experimenter; quenched simulations were readily identifiable because of a relative enrichment of poorly packed and extended structures at low temperatures. For some of the larger targets, simulation times exceeded 48h/replica. The previously used selection protocol, also used here, entails clustering the structures by their pairwise difference in Q (the number of contacts). Other intuitive metrics were used, like the number of hydrophobe-hydrophobe contacts (a quantity to be maximized), and the number of hydrophobe-polar contacts (to be minimized). Structures that simultaneously satisfied all criteria were usually selected for submission.

A new selection criteria was used this year, casually dubbed 'frustratometry.' [8,9] For a given pair of residues in the structure, we calculate the change in apparent stability from a perturbation to the local density of contacts surrounding the interacting pair. This is then repeated for all interacting pairs of residues. Comparing the actual stability with the distributions of stabilities over varying densities, a Z-score is assigned to establish whether the pair is typical for the protein, is relatively destabilized, or is relatively stabilized. This Z-score is called the configurational frustration index. Adding up the numbers of stabilized and destabilized contacts for that structure gives two additional summary statistics: the number of minimally frustrated residues, and the number of maximally frustrated residues. The latter quantity, when minimized, identifies candidates that would have been missed in previous competitions.

1. Friedrichs.M.S.,Wolynes,P.G. (1989) Science 246:371
2. Friedrichs,M.S.,Wolynes,P.G. (1990) Tetrahedron Computer Methodology 3:175-190
3. Koretke,K.K.,Luthery-Schulten,Z.,Wolynes,P.G. (1996) Protein Science 5:1043
4. Eastwood,M.P.,Hardin,C.,Luthey-Schulten,Z.,Wolynes,P.G. (2001) IBM J. Res. & Dev. 45:475
5. Hardin,C.,Eastwood,M.P.,Prentiss,M.C.,Luthery-Schulten,Z.,Wolynes,P.G. (2003) Proc. Natl. Acad. Sci. USA 100:1679
6. Papoian,G.A.,Ulander,J.,Eastwood,M.P.,Wolynes,P.G. (2004) Proc. Natl. Acad. Sci. USA 101:3352
7. Cole,C.,Barber,J.D.,Barton,G.J. (2008) Nucleic Acids Res. 36(2): W197
8. Ferreiro,D.U.,Hegler,J.A.,Komives,E.K.,Wolynes,P.G. (2007) Proc. Natl. Acad. Sci. USA 104:19819
9. http://www.frustratometer.tk/

YASARA

The YASARA homology modeling module V2.0 with improved alignments, oligomerization and a new hires refinement force field

E. Krieger
CMBI 260, NCMLS, Radboud University Nijmegen Medical Center,
PO Box 9101, 6500 HB Nijmegen, the Netherlands, elmar@cmbi.ru.nl

Like in CASP8, the 'YASARA Structure' server (*www.yasara.org/ homologymodeling*) submitted predictions for those targets that could be built reliably using known template structures. CASP8 evaluation identified alignment accuracy as the main bottleneck, which has therefore been the development focus, while hires refinement needed less attention[1].

Methods

The YASARA homology modeling module employed the following recipe: PSI-BLAST[2] was run against Uniprot to build a target PSSM, which was then used to find the five closest templates in the PDB. For each template, a profile was created using related sequences and related structures, which were obtained from a FatCat[3] based all-against-all comparison hosted at the RCSB (3D similarity tab). For each of the five templates, up to five stochastic profile-profile alignments were created[4] using SSALIGN scoring matrices[5]. And for each of the maximally 25 template/alignment combinations, a 3D model was built using loop conformations extracted from the PDB and an improved SCWRL side-chain placement algorithm[6]. After an extended refinement minimization[1], the models were ranked by quality Z-score, and the top five were submitted. The following **special features** were handled automatically: inclusion of ligands in the model (as long as they interact well and stabilize the structure), automatic oligomerization to capture stabilizing effects of quaternary structure and pH-dependent hydrogen bonding networks that include ligands to aid hires refinement.

Results

The recipe above yielded models with reliable quality scores for 70 of the 129 CASP9 targets. The server was deliberately configured not to submit models that were considered incorrect and is therefore incompatible with a ranking scheme that simply sums up GDT_TS values over all targets and includes fold recognition and de novo folding. The current focus is just on high-resolution homology modeling needed e.g. for drug design.

Availability

The homology modeling module described here is available as part of YASARA Structure from www.yasara.org/products.

1. Krieger E, Joo K, Lee J, Lee J, Raman S, Thompson J, Tyka M, Baker D, Karplus K (2009). Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in CASP8. *Proteins 77 Suppl* **9**, 114-122
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. Yuzhen Ye & Adam Godzik (2003). Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics* 19 suppl. 2. ii246-ii255.
4. Mueckstein U, Hofacker IL and Stadler PF (2002). Stochastic pairwise alignments. *Bioinformatics* 18 Sup2, 153-160
5. Qiu J and Elber R (2006). SSALN: An alignment algorithm using structure-dependent substitution matrices and gap penalties learned from structurally aligned protein pairs. *Proteins* 62, 881-891
6. Canutescu AA, Shelenkov AA and Dunbrack RL Jr. (2003). A graph-theory algorithm for rapid protein side-chain prediction.*Protein Sci.* **12**, 2001-2014.

Yuan-Chen-Kihara

Template-based protein structure prediction by SUPRB threading method

Chao Yuan[1], Hao Chen[1] and Daisuke Kihara[1, 2]
[1] – Department of Biological Sciences, [2] – Department of Computer Science
College of Science, Purdue University, West Lafayette, IN, 47907, USA
dkihara@purdue.edu

We submitted predictions in the tertiary structure prediction (TS) category. In total of 610 models are submitted, five models each for the 62 targets (2 of the targets have been later cancelled). We used SUPRB (threading with SUboptimal alignment-based PRoBabilistic residue contact information), a threading method which is recently developed in our group.

Methods

The SUPRB threading algorithm [1] employs a sequence-structure compatibility score which linearly combines five different scoring terms to evaluate fitness of a query sequence to template structures. The five terms are a sequence profile term, a secondary structure term, a solvent accessibility term, a main chain angle potential term, and an amino acid contact potential term. The target-template alignments are computed by the dynamic programming (DP) algorithm. It is known that the DP algorithm is not able to compute the optimal alignment for a pairwise residue contact potential. To accommodate this problem, SUPRB computes alignments iteratively, by first computing alignments without the contact term and using the contact term in the subsequent rounds. For each round, both the optimal and over hundred of suboptimal alignments are computed [2]. The contact term is used either to re-rank the optimal and suboptimal alignments computed without the contact term or to update alignments by adding the contact term in several rounds. In the latter strategy, the contact term is handled in a probabilistic fashion by using the suboptimal alignments. Namely, residue contacts inferred from each suboptimal alignment in the previous round are counted and the fraction of suboptimal alignments inferring each contact is used as a weighting term for the score of each residue contact pair. On a benchmark dataset, the probabilistic handling of the contacts were shown to outperform the partly thawed approach [3], which only uses the optimal alignment in defining residue contacts [1]. Using suboptimal alignments is also shown to improve homology modeling by MODELLER [1].

For CASP9, SUPRB was run against a representative template structure dataset consisting of 10926 structures and ranked templates according to the optimal alignment score normalized by sequence length. Then, the tertiary structures of the top scoring templates are built by MODELLER by feeding the optimal alignments and four top scoring suboptimal alignments.

Results

Among the first 19 targets whose native structures were posted on the CASP9 official website, SUPRB predicted 7 targets within an RMSD of 6Å.

Availability

SUPRB is available at http://www.kiharalab.org/suprb/ for download.

Acknowledgements

This work is supported by NSF (EF0850009, IIS0915801) and NIH (GM075004). Other support from NSF is also acknowledged (DMS80568).

1. Chen H, Kihara D. (2010) Effect of using suboptimal alignments in template-based protein structure prediction. Proteins, in press.
2. Chen H, & Kihara D. (2008). Estimating quality of template-based protein structure models by alignment stability. Proteins 71, 1255-1274.
3. Skolnick J, Kihara D. (2001) Defrosting the frozen approximation: PROSPECTOR-a new approach to threading. Proteins, 42: 319-331.

Zhang
Zhang-Server
QUARK

Automated structure predictions by I-TASSER and QUARK pipelines

Yang Zhang, Dong Xu, Jian Zhang, Ambrish Roy, Jinrui Xu
Center for Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw Ave, Ann Arbor, MI 48109
zhng@umich.edu

The procedures we used for the human (as "Zhang") and server (as "Zhang-Server" and "QUARK") predictions are depicted in Figure 1. Methods used by "Zhang" and "Zhang-Server" are based on I-TASSER and essentially the same, except for that the human prediction exploited the templates in CASP9 Server Section while Zhang-Server used our in-house threading programs. QUARK is a new pipeline developed mainly for *ab initio* protein structure assembly[1; 2]. All the procedures are fully automated.

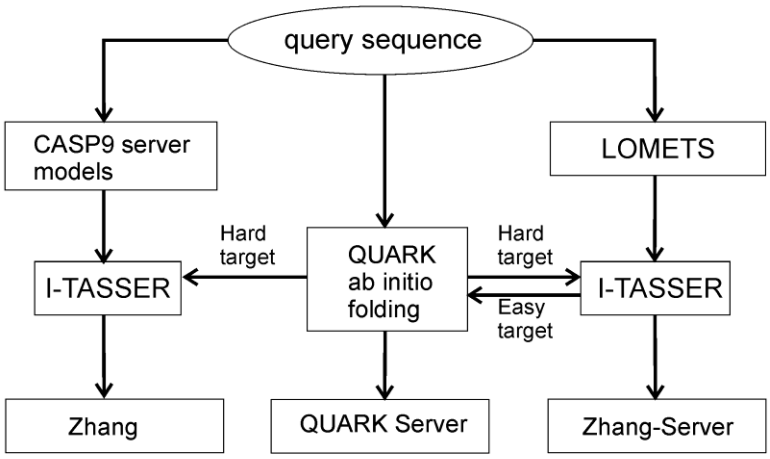


Figure 1. Flowchart for automated structure modeling generated for "Zhang", "Zhang-Server", and "QUARK" in CASP9.

Compared to our previous prediction procedures[3; 4], the major new developments in CASP9 are QUARK for *ab initio* protein folding[1; 2], and FG-MD for protein structure refinements[5].

QUARK *ab initio* structure assembly. For a given sequence, QUARK first generates small structural fragments with length [1, 20] by gapless threading through the PDB library[6]. The fragments are ranked based on a composite scoring function consisting of sequence and structure profiles, and predicted secondary structure and torsion angles. Top 200 fragments at each position are exploited to assemble the 3D model of the target sequence by replica-exchange Monte Carlo simulations, under the guidance of an atomic knowledge-based potential, assisted with the distance profiles collected from the fragment library[1]. No global template information is used in QUARK simulations.

FG-MD structure refinement. The fragment-guided molecular dynamics (FG-MD) simulations are implemented by LAMMPS (lammps.sandia.gov)[7], with the force field consisting of terms from Amber99[8], C-alpha repulsive potential, H-bonding network, and structural fragments searched by TM-

align using the initial model as probe through the PDB library. The contact and distance restraints are collected from the top 20 TM-align fragments which are used to constrain the MD simulations[5].

The prediction pipelines include three general steps: template identification, structure reassembly, model selection and refinements.

Template identification. The target sequences are first threaded through non-redundant PDB structure libraries for identifying appropriate template alignments by LOMETS[9]. In human prediction, we include additionally the models generated by other groups in the Server Section into the template pool. Having more threading templates from the Server Section is the only source of differences between Zhang and Zhang-Server predictions. The degree of structural consensus of multiple templates, assessed by the average TM-score, is used to categorize the targets into "easy" or "hard".

Template-based or *ab initio* Structure assembly. In I-TASSER, continuous fragments excised from the threading templates are exploited to assemble full-length models[10; 11; 12] with unaligned loop regions built by *ab initio* modeling[13]. The I-TASSER potential includes four components: (1) general knowledge-based statistics terms from the PDB (C-alpha/side-chain correlations[13], H-bond[14] and hydrophobicity[15]); (2) spatial restraints from threading templates[9]; (3) sequence-based C-alpha contact predictions by SVMSEQ[16]; (4) distance map from segmental threading[17]. For hard target, I-TASSER also use restraints from models generated by QUARK. In QUARK Server, for "easy" targets the template restraints from I-TASSER models are used to guide the QUARK simulations while for "hard" targets models are generated by the *ab initio* folding without template restraints (Figure 1).

Model selection and refinements. The structures in low-temperature replicas of I-TASSER and QUARK simulations are clustered by SPICKER[18]. The atomic models are constructed by REMO[19] from the cluster centroids by the optimization of the hydrogen-bonding network which is predicted by secondary structure assignments and the 3D backbone model. Finally, all the models are submitted to FG-MD[5] for structure refinement before submission, with the purpose of improving local geometry and H-bonding, and reducing steric clashes of the models.

The on-line I-TASSER and QUARK servers are available, respectively, at:
http://zhanglab.ccmb.med.umich.edu/I-TASSER
http://zhanglab.ccmb.med.umich.edu/QUARK.

1. Xu, D. & Zhang, Y. (2010). QUARK ab initio protein structure prediction I: Method developments. Submitted.
2. Xu, D. & Zhang, Y. (2010). QUARK ab initio protein structure prediction II: results of benchmark and blind tests. Submitted.
3. Zhang, Y. (2007). Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins* **69**, 108-117.
4. Zhang, Y. (2009). I-TASSER: Fully automated protein structure prediction in CASP8. *Proteins* **77**, 100-113.
5. Zhang, J. & Zhang, Y. (2010). High-resolution protein structure refinement using fragment guided molecular dynamics. Submitted.
6. Xu, D. & Zhang, Y. (2010). What is the optimal fragment length for ab initio protein structure assembly? , Submitted.

7. Plimpton, S. (1995). Fast Parallel Algorithms for Short-Range Molecular-Dynamics. *Journal of Computational Physics* **117**, 1-19.
8. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W. & Kollman, P. A. (1996). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules (vol 117, pg 5179, 1995). *Journal of the American Chemical Society* **118**, 2309-2309.
9. Wu, S. T. & Zhang, Y. (2007). LOMETS: A local meta-threading-server for protein structure prediction. *Nucl. Acids. Res.* **35**, 3375-3382.
10. Zhang, Y. & Skolnick, J. (2004). Automated structure prediction of weakly homologous proteins on a genomic scale. *Proc. Natl. Acad. Sci. USA* **101**, 7594-7599.
11. Wu, S., Skolnick, J. & Zhang, Y. (2007). Ab initio modeling of small proteins by iterative TASSER simulations. *BMC Biol* **5**, 17.
12. Roy, A., Kucukural, A. & Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc* **5**, 725-38.
13. Zhang, Y., Kolinski, A. & Skolnick, J. (2003). TOUCHSTONE II: A new approach to ab initio protein structure prediction. *Biophys. J.* **85**, 1145-1164.
14. 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. USA* **103**, 2605-10.
15. Chen, H. & Zhou, H. X. (2005). Prediction of solvent accessibility and sites of deleterious mutations from protein sequence. *Nucleic Acids Res* **33**, 3193-9.
16. Wu, S. & Zhang, Y. (2008). A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. *Bioinformatics* **24**, 924-31.
17. Wu, S. & Zhang, Y. (2010). Recognizing protein substructure similarity using segmental threading. *Structure* **18**, 858-67.
18. Zhang, Y. & Skolnick, J. (2004). SPICKER: A clustering approach to identify near-native protein holds. *J Comput Chem* **25**, 865-71.
19. Li, Y. & Zhang, Y. (2009). REMO: A new protocol to refine full atomic protein models from C-alpha traces by optimizing hydrogen-bonding networks. *Proteins* **76**, 665-76.

Zhang_Ab_Initio

Ab initio protein structure prediction by QUARK combined with human interventions

Dong Xu and Yang Zhang
Center for Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw Ave, Ann Arbor, MI 48109
zhng@umich.edu

QUARK is a newly developed program for *ab intio* protein structure assembly[1; 2]. Although models of all the human targets were submitted, Zhang_Ab_Initio mainly focused on modeling of the hard, new fold targets. The targets assigned by LOMETS[3] as "hard" are modeled by QUARK *ab initio* simulations while the models for the "easy" targets were selected from the predictions in Server Section. Several server targets were also submitted if they are judged by LOMETS as "hard" targets (without templates in the PDB library). The human-intervention was mainly on the server model selections and domain split of multiple-domain proteins.

QUARK contains four steps. First, it runs PSI-BLAST[4] and PSIPRED[5] to obtain sequence profile and secondary structure prediction. Solvent accessibility, real-value torsion angles, beta-turns are predicted by neural network training. Structural fragments with lengths in the range of [1, 20] are generated by gapless threading which are ranked based on a composite scoring function of the obtained information[6]. The second step is to assemble the fragments into 3D models by replica-exchange Monte Carlo simulation. The QUARK potential consists of multiple knowledge-based energy terms and the movements include fragment replacements and free backbone heavy atom moves. Third, decoys of the 10 lowest temperatures are clustered by the SPICKER program[7] and 10 models closest to the cluster centroids are returned. Fourth, full atomic models are built from the backbone models with side-chains built by SCWRL4[8] and refined using FG-MD[9]. Five models are finally selected by the model quality assessment programs.

Multiple domain proteins and the domain boundaries were defined by LOMETS alignments and human visualization. For proteins with multiple hard domains, QUARK simulations were conducted on individual domains, which were then assembled into full-length model which kept the core region rigid with the linker regions constructed by QUARK *ab initio* prediction[10]. For proteins having easy template-based domains, QUARK was conducted on the hard domain while keeping the other easy domain structure fixed.

1. Xu, D. & Zhang, Y. (2010). QUARK ab initio protein structure prediction I: Method developments. Submitted.
2. Xu, D. & Zhang, Y. (2010). QUARK ab initio protein structure prediction II: results of benchmark and blind tests. Submitted.
3. Wu, S. T. & Zhang, Y. (2007). LOMETS: A local meta-threading-server for protein structure prediction. *Nucl. Acids. Res.* **35**, 3375-3382.
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**, 3389-402.
5. Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195-202.
6. Xu, D. & Zhang, Y. (2010). What is the optimal fragment length for ab initio protein structure assembly? , Submitted.

7. Zhang, Y. & Skolnick, J. (2004). SPICKER: A clustering approach to identify near-native protein folds. *J Comput Chem* **25**, 865-71.
8. Krivov, G. G., Shapovalov, M. V. & Dunbrack, R. L., Jr. (2009). Improved prediction of protein side-chain conformations with SCWRL4. *Proteins* **77**, 778-95.
9. Zhang, J. & Zhang, Y. (2010). High-resolution protein structure refinement using fragment guided molecular dynamics. Submitted.
10. Xu, D. & Zhang, Y. (2010). Free and restricted multipledomain protein structure prediction by QUARK simulation. In preparation.

Zhang_FUNCTION
I-TASSER_FUNCTION

Binding site predictions using COFACTER algorithm

Ambrish Roy and Yang Zhang
Center for Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw Ave, Ann Arbor, MI 48109, USA
zhng@umich.edu

Binding site predictions for our human (as "Zhang_FUNCTION") and server (as "I-TASSER_FUNCTION") predictions are both based on COFACTOR, a recently developed structure-based binding site prediction program[1]. While the human predictions used multiple 3D models predicted by other servers in CASP9 Server Section, I-TASSER_FUNCTION exploited only the 3D models predicted by the "Zhang server". The COFACTER procedure includes two stages: (a) Functional analog identification; (b) binding site recognition by local motif match.

Functional analog identification. Structure and functional analogs of query proteins are identified by matching the 3D models through a template library of known binding sites using the structural alignment program TM-align[2]. All the template proteins in the library with similar folds, i.e. having a TM-score >0.5[3], were considered as potential candidates of function analogs, and used as an input for the next step of binding site refinement by local structural motif match.

Binding site recognition. Binding site refinement by COFACTOR involves four steps:

a. *Generation of candidate binding site motifs in query*: Conserved residues in query, which have the same identity to the binding site residues in the templates, are identified based on their Jensen–Shannon divergence score[4] and are marked as potential binding site locations. The structures of all combined sets of marked residues are excised from the predicted model and are used as candidate binding site motifs.
b. *Superposition of candidate binding site motifs onto template binding site*: These local 3D candidate motifs of query protein are superimposed onto template's binding site residues. For each residue *i*, the coordinates of two neighboring residues, i.e. *i*-1 and *i*+1th residues are also used to increase the reliability of structural superimposition. The rotation and translation matrix acquired from this superimposition is used for superposing the complete structure of query and template proteins. A putative binding site region in query's predicted structure is then defined using a sphere of radius *r*, where *r* is the maximum distance of the template residues from the geometric center of template binding site.
c. *Alignment of putative and template binding site*: The best alignment between the query and template binding sites, i.e. the region defined within the sphere of radius *r*, is identified using an iterative Needleman-Wunsch dynamic programming[5], where the score for aligning *i*th residue in query and *j*th residue in template is given by the sum of BLOSUM62 residue similarity and reciprocal distance between the residues. For each alignment, the final raw alignment score is calculated as the sum of structure and sequence match over all the aligned residue pairs, normalized by the number of residues present in the template's binding site.
d. *Identification of binding site*: Step (a) to (c) is repeated for all candidate binding site motifs. The region which gives the best binding site score (BS-score) is selected as the identified binding site in the query; the residues aligned with known binding site residues in the template as binding site residues.

Availability

The COFACTER algorithm is implemented on the I-TASSER server (http://zhanglab.ccmb.med.umich.edu/I-TASSER) starting from a sequence, and http://zhanglab.ccmb.med.umich.edu/PSFloger, starting from a 3D structure.

1. Roy, A. & Zhang, Y. (2010). COFACTOR: protein binding site recognitions by global structure match and local geometry refinement. Submitted.
2. Zhang, Y. & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* **33**, 2302-9.
3. Xu, J. & Zhang, Y. (2010). How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **26**, 889-95.
4. Capra, J. A. & Singh, M. (2007). Predicting functionally important residues from sequence conservation. *Bioinformatics* **23**, 1875-82.
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.

Zhang-Refinement

High-resolution protein structure refinement by FG-MD

Jian Zhang, Yang Zhang
Center for Computational Medicine and Bioinformatics, University of Michigan,
100 Washtenaw Ave, Ann Arbor, MI 48109
zhng@umich.edu

Fragment-guided molecular dynamics (FG-MD) simulations[1] were performed for all the refinement targets in CASP9. Given an initial protein 3D model, we first searched through our non-redundant PDB structure library for identifying experimental structures which are closest to the initial model. TM-align[2] was used for searching the PDB library and the experimental structures were ranked based on their TM-score[3] to the initial model. Top 20 experimental structures were selected for generating C-alpha fragments, which are the structurally aligned region of the experimental structure with the initial model. For each experiment structure, a set of C-alpha distance restraints were taken from these fragments and added to our composite potential, which included Amber99 force field[4], C-alpha repulsive potential, C-alpha contact distance restraint and statistical hydrogen bonding potential. Simulated annealing molecular dynamics was used to sample the local energy minima, which was implemented in the molecular dynamics code LAMMPS (http://lammps.sandia.gov/)[5]. The refined models were selected, from the top 20 generated decoys, based on the number of backbone hydrogen bonds and number of steric clashes. Statistical hydrogen bonding potential was used to enforce the hydrogen bonding (HB) network, which is a list of HB donor–acceptor atom pairs. HB network is constructed based on the predicted secondary structure distribution and the global structure of the initial model[6]. The procedure is fully automated and the running time for each refinement target is less than 1 hour. The on-line FG-MD server is available at http://zhanglab.ccmb.med.umich.edu/FG-MD.

1. Zhang, J. & Zhang, Y. (2010). High-resolution protein structure refinement using fragment guided molecular dynamics. Submitted.
2. Zhang, Y. & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research* **33**, 2302-9.
3. Zhang, Y. & Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702-710.
4. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W. & Kollman, P. A. (1996). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules (vol 117, pg 5179, 1995). *Journal of the American Chemical Society* **118**, 2309-2309.
5. Plimpton, S. (1995). Fast Parallel Algorithms for Short-Range Molecular-Dynamics. *Journal of Computational Physics* **117**, 1-19.
6. Li, Y. & Zhang, Y. (2009). REMO: A new protocol to refine full atomic protein models from C-alpha traces by optimizing hydrogen-bonding networks. *Proteins-Structure Function and Bioinformatics* **76**, 665-76.

ZHOU-SPARKS-M

Using Neural Networks to Aid a Human in Predicting Protein Structure

E. Faraggi[1,2], Y. Yang[1,2] and Y. Zhou[1,2]

[1] – Center for Computational Biology and Bioinformatics, School of Medicine, Indiana University, Indianapolis, Indiana, USA, [2] - School of Informatics, Indiana University Purdue University, Indianapolis, Indiana, USA

efaraggi@iupui.edu

Predicting a protein's structure from its amino-acid sequence can be a challenging endeavor. If the sequence in question is homologous to proteins with known structure the problem is relatively easy. However, if no related structures are known or found the problem is considerably more difficult. In this work we show how a neural network can be used to aid in the ranking of structural templates and other models for the human prediction part of the CASP meeting.

Methods

Given an amino acid sequence of a protein fragment (chain), the starting position for modeling its three dimensional structure is its sequence alignments to known structural templates using the SPARKS-X method.[1] Out of these alignments a set of features was constructed to account for the similarity between the query and template sequences, and the appropriateness of using a given known structure as a template model. Generally, these inputs are grouped into features dependent on query, on template, and on query-template alignment. A total of 314 input features were generated for each query-template pair.

A GEneral Neural Network (GENN) which was locally developed was used to train a set of weights to predict the TM-score[2] between each of the query-template pairs. The general architecture chosen was of a two hidden layer network with 51 nodes per hidden layer, and two output nodes that were simultaneously trained and their output averaged to generate the network's TM-score prediction. Due to time constraints no optimization was performed on the architecture of the neural network. A dataset of non-homologous proteins with a sequence similarity cutoff of 30% was used to train the neural network. The template library was based on a 40% sequence similarity cutoff of PDB sequences. While training we removed any query-template pairs with a sequence identity greater than 30%.

Training query sequences were divided into three groups. The hard group was composed of sequences whose top-one z-score, as calculated by SPARKS-X, was smaller than 8. Typically, template based structure prediction of such proteins has limited success: reasonable template structures may not exist and if they do exist their identification is difficult. For this group 14 separate sets of weights were trained to predict the TM-score and the final prediction was taken as their average. The medium group was chosen as those queries with top-one z-scores smaller than 15. This group was further split specifically separating queries with z-score between 8 and 15. This group typically has reasonable template models but their identification by the z-score method is not always correct. Sixteen sets of weights were trained for this group. Finally, cases with z-scores between 15 and 20 were treated as easy cases and 6 neural networks were trained on this group. Typically, queries with such high z-score can be modeled rather successfully, however, mistakes can be found.

For real-world prediction on CASP9 queries, SPARKS-X was first run. From its alignments and raw-scores, input features were constructed and were fed into the set of weights best corresponding to its top-one template z-score. Predicted TM-scores are then ordered, and the maximum is chosen as top-one prediction. Human intervention occurred when the alignments were poor; the resulting structures were

significantly non-compact; or if significantly more "pleasing" structure models were found by other groups.

Results

Initial training and testing of the proposed approach were done on the SCOP database. For hard cases, the probability of identifying the correct top-one model (based on TM-score) increased from 12% using z-score to 15% using GENN. The probability of identifying the top-one TM-score in the top-five ranked templates increased from 38% to 41%. The average TM-score between top-one ranked model and native structure increased from 0.396 to 0.403. For hard and medium cases, top-one increased from 53% to 56%, while top-five increased from 78% to 79%. The overall average TM-score increased from 0.636 to 0.641.

Testing was also performed during the training of the weight on the SPARKS-X template database. In this case top-one accuracy for hard queries increased from 52.8% to 54.2%, top-one accuracy for medium cases increased from 83.6% to 85.2%. The Pearson correlation coefficient between predicted native-to-model TM-scores and actual native-to-model TM-scores were approximately 0.9. Based on these results a server was constructed and used to predict structure for the CASP8 targets. The overall average TM-score for this test set increased from 0.55 to 0.58 with the use of GENN.

Discussion/Conclusions

Model templates for protein tertiary structure prediction were re-ranked using a neural network trained to predict the TM-score between query-template pairs. High correlation was found for these predictions and they were found to be significantly useful in discriminating top structural templates. It was also found that the most significant improvement, resulting from the use of the proposed neural network, arises for those cases with low z-score. That is, those hardest to predict cases.

Availability

World Wide Web implementation of the procedures outlined above will be available through the SPARKS-X package that is currently under development.

1. See SPARKS-X abstract in current publication. The latest published version of SPARKS can be found in W. Zhang, S. Liu, Y. Zhou (2008) *SP5: Improving protein fold recognition by using predicted torsion angles and profile-based gap penalty,* PLoS ONE **3**, e2325
2. Y. Zhang, J. Skolnick (2004) *Scoring function for automated assessment of protein structure template quality*, Proteins **57**, 702-710.

ZHOU-SPARKS-X

SPARKS-X: Improving the single fold-recognition technique by employing statistical error potentials.

Y. Yang, E. Faraggi and Y. Zhou*
Center for Computational Biology and Bioinformatics, School of Medicine, Indiana University, Indianapolis, Indiana, USA, and School of Informatics, Indiana University Purdue University, Indianapolis, Indiana, USA
yueyang@iupui.edu and yqzhou@iupui.edu

Fold recognition refers to recognizing the structural fold of a protein from its sequence. In recent CASP tests, although the best structure prediction servers involve some post-treatment of predicted models, the prediction quality of these methods is mostly determined by the quality of the template recognized. A series of successful single fold-recognition methods were developed in our group (SPARKS, SP2, SP3, SP4, SP5) [1] that use both sequence profiles from multiple sequence alignment, and structure profiles, including secondary structure (SS), solvent accessible surface area (ASA) and main-chain torsion angles (φ/ψ). Here, we further improve the method by employing statistical error potentials to estimate the agreement between the native template structure and improved predicted structural properties of the query sequence such as SS, φ/ψ, and ASA.

Methods

Alignment Score:

The alignment score for aligning template position *i* with the query position *j* is

$$s(i,j) = -\frac{1}{200}\sum_{k=1}^{20}\left[F_t(i,k)M_q(j,k) + M_t(i,k)F_q(j,k)\right] + w_{ss}E_{ss}(ss_t(i)\left|ss_q(j), C_{ss,q}(j)\right.)$$
$$+ w_\tau \sum_{k=1}^{2} E_\tau(\Delta\tau^k(i,j)\left|C_{\tau^k,q}(j)\right.) + w_A E_A(\Delta A_{ij}\left|R_q(j)\right.) + w_{\text{shift}}$$

with three weight parameters ($w_{ss}w_\tau w$with three weight parameters (w_{ss},w_τ,w_A) for secondary structure, torsion angles and solvent accessible area and one constant shift w_{shift}. The first term in the equation is the profile-profile comparison between the sequence profile from the query sequence and that from the template sequence. $F_t(i,k)$and $F_q(j,k)$are the frequency from the sequence-derived frequency profile of the template sequence and that of the query sequence, respectively; $M_t(i,k)$ and $M_q(j,k)$ are the sequence-derived log-odds profile of the template sequence and query sequence, respectively. Unlike SP3, SP4 and SP5, one sequence profile from the structure of the template is not employed because its effect is no longer significant with improved prediction of structure properties. The second term measures the difference between the native secondary structure assignment of the template given by the program *DSSP* and the predicted secondary structure of the query protein by the server *SPINE-X*[2]. $C_{ss,p}$ is the predicted confidence score also by the server. The function is calculated from the statistics on the dataset of 2479 proteins, which was used to train the SPINE-X server, according to the equation:

$$E_{ss}(ss_0\left|ss_p, C_{ss,p}\right.) = -\ln\left(\frac{P(ss_0\left|ss_p, C_{ss,p}\right.)}{P(ss_0)}\right)$$

where ss_0 is the actual secondary structure, ss_p and $C_{ss,p}$ are the predicted secondary structure and confidence score, respectively. Here, the secondary structure has three standard states, and the confidence score is evenly divided into eight discrete states. Similarly, we have developed terms for torsion angles $E_\tau(\Delta\tau\left|C_\tau\right.)$ and an amino acid type dependent term for solvent accessible surface area $E_A(\Delta A\left|R\right.)$.

Parameter Training and Template Ranking

The Smith-Waterman local alignment algorithm is used to optimize the score that matches the query profiles with template profiles. All four weights parameters and two gap penalty parameters (gap opening g_o and gap extension g_e) were trained on the Prosup Benchmark. The parameters were trained using the Powell method by many repeats from different random seeds. The final parameters used are w_{ss}=0.95, w_τ=0.75, w_A=2.37, g_o=12.4, g_e=0.66, w_{shift}=-1.52.

The templates are ranked by the greater one of two z-scores, which is calculated based on the raw alignment score normalized by the full alignment length and the non-end-gap alignment length, respectively. The score is normalized by S/L^{α}. Based on tests and trials α is set to 3/4.

Template Library

An automatically updated template library is used for the threading. When a new protein is inputted to the library, it is first divided into domains according to the "Author" parameters in DDOMAIN[3]. The divided domains are compared to all existing domains in the library. If the sequence identity is less than 40%, or the TM-score (by TM-align) is smaller than 0.5, the new domain and its corresponding chain were included in the library. By this way, the library had 31750 templates by Jul 2010.

Model building

The model is built by modeller9v7 using the alignment generated by SPARKS-X. When there are gaps of more than 30 residues in the termini, the procedure will be reused to build a separate model for the missing part. Subsequently, a refinement program was used to link the models of different parts of the query sequence and remove clashes by using the DFIRE potential function[4].

Availability

The server is available on http://sparks.informatics.iupui.edu/sparks-x

1. The last SPARKS version: Zhang,W., Liu,S. & Zhou,Y. (2008). SP5: improving protein fold recognition by using torsion angle profiles and profile-based gap penalty model. PLoS One 3, e2325.
2. Faraggi,E., Yang,Y., Zhang,S. & Zhou,Y. (2009). Predicting continuous local structure and the effect of its substitution for secondary structure in fragment-free protein structure prediction. Structure 17, 1515-27.
3. Zhou,H., Xue,B. & Zhou,Y. (2007). DDOMAIN: Dividing structures into domains using a normalized domain-domain interaction profile. Protein Sci 16, 947-55.
4. Yang,Y. & Zhou,Y. (2008). Specific interactions for ab initio folding of protein terminal regions with secondary structures. Proteins 72, 793-803.

ZHOU-SPINE-D

Intrinsic disorder prediction using neural networks

T. Zhang[1,2], E. Faraggi[1,2] and Y. Zhou[1,2,*]
[1] – *School of Informatics, Indiana University Purdue University, Indianapolis, Indiana 46202,* [2] – *Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, Indiana 46202*
yqzhou@iupui.edu

Intrinsically disordered proteins and intrinsically disordered regions in proteins do not fold into stable three-dimensional structures under general physiological conditions. Although lacking specific structures, disorder regions play crucial functional roles in many biological processes, such as transcriptional regulation, translation and cellular signal transduction; they are also shown to be prevalent in various human diseases including cancer, cardiovascular disease, and genetic diseases, to name a few. Disordered regions also play an important theoretical role, containing information about the relationship between sequence and structure (or lack of it). Due to their functional and theoretical importance, a way to reliably identify disordered regions would be highly beneficial. Experimental techniques for disorder detection are costly and time-consuming, sometimes even impossible. Thus, many computational algorithms have been developed for predicting disordered regions. Here, we introduce a new predictor, SPINE-D. It incorporates a number of sequence-based features and provides accurate predictions of disordered regions.

Methods

A large set of proteins was prepared for training and testing SPINE-D. It includes all X-ray-determined structures in PDB having residues without atomic coordinates, as well as the fully disordered proteins released in the Disprot[1] database. All the protein chains were then filtered by blastclust[2] to ensure that the pairwise sequence identity is below 25%. This resulted in a set of 4229 protein chains, named DP4229. The DP4229 dataset was divided into two subsets: the DP3000 dataset (3000 chains) for designing our predictor; and the DP1229 dataset (1229 chains) for blind test.

We used a two-hidden-layer neural network with a filter predictor for smoothing the predictions, a hyperbolic activation function and guided learning technique developed for Real-SPINE 3.0[3]. Considering that the primary sequences in long and short disordered regions are dissimilar, we sorted residues into 3 classes: ordered residues, residues in short disordered regions (<= 30 residues) and residues in long disordered regions (>30 residues). Later, predictions for the latter two classes were combined to yield the final disorder predictions. To reduce random prediction errors caused by the randomly selected initial weights, we trained five independent predictors and the final prediction is based on their consensus.

The input nodes used residue-level and window-level information, as well as one terminal tag. The residue-level information includes: (a) seven representative physical parameters identified by Meiler *et al* [4]; (b) a PSSM vector derived from the PSI-BLAST[2] profiles; (c) predicted secondary structure and solvent accessibility from SPINE-X[5]; and (d) predicted torsion-angle fluctuation[6]. A sliding window of size 21, centered on the current residue, was introduced to include the information of its neighboring residues. In terms of window-level information, we considered the current residue plus 15 residues on either side, and calculated: (a) amino acid composition; (b) local compositional complexity[7]; (c) predicted secondary structure content. The terminal tag marked residues on both N- and C-termini.

We applied a cost matrix to accommodate for the imbalanced populations of disordered and ordered residues, i.e., higher weights were given to the minority (disordered) residues. More specifically, we made duplicates for disordered samples during the training process.

The window size and parameters were optimized for the highest AUC (area under the ROC curve) value. Once the neural network output the probability for order/disorder predictions, we picked up the threshold that led to the highest S_w score[8] to yield binary predictions. All optimizations were done on the training sets, accuracy results quoted bellow are for the test sets.

Results

Ten-fold cross validation test was performed on the DP3000 dataset. Our method achieved an AUC of 0.858 and a S_w of 0.574. Similar results, with an AUC of 0.859 and a S_w of 0.572, were observed when we trained our prediction model on the entire DP3000 dataset and then tested on the blind DP1229 dataset.

Our final predictor SPINE-D was trained on the DP4229 dataset. We tested SPINE-D on the CASP8 dataset. The AUC was 0.908, and S_w equaled 0.693.

Availability

SPINE-D is available at http://sparks.informatics.iupui.edu/SPINE-D/index.html.

1. Sickmeier M., Hamilton J.A., LeGall T., Vacic V., Cortese M.S., Tantos A., Szabo B., Tompa P., Chen J., Uversky V.N., Obradovic Z., Dunker AK. (2007). DisProt: the Database of Disordered Proteins. *Nucleic Acids Res.* **35** (Database issue): D786-93.
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. Faraggi E., Xue B., Zhou Y. (2009) Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two-layer neural network. *Proteins* **74**: 847–856.
4. Meiler J., Muller M., Zeidler A., Schmaschke F. (2001) Generation and evaluation of dimension reduced amino acid parameter representations by artificial neural networks. *J. Mol. Model.* **7**: 360–369.
5. Faraggi E., Yang Y., Zhang S., Zhou Y. (2009) Predicting continuous local structure and the effect of its substitution for secondary structure in fragment-free protein structure prediction. *Structure* **17**: 1515–1527.
6. Zhang T., Faraggi E., Zhou Y. (2010) Fluctuations of backbone torsion angles obtained from NMR-determined structures and their prediction. *Proteins*, in press.
7. Wootton J.C. (1994) Sequences with 'unusual' amino acid compositions. *Curr. Opin. Struct. Biol.* **4**: 413–421.
8. Jin Y., Dunbrack R.L. Jr. (2005) Assessment of disorder predictions in CASP6. *Proteins*. **61** (Suppl 7) :167-175.

ZHOU-SPINE-DM

Meta server approach for intrinsic disorder prediction

T. Zhang[1,2], E. Faraggi[1,2] and Y. Zhou[1,2,*]
[1] – School of Informatics, Indiana University Purdue University, Indianapolis, Indiana 46202, [2] – Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, Indiana 46202
yqzhou@iupui.edu

Intrinsically disordered proteins and intrinsically disordered regions in proteins do not fold into stable three-dimensional structures under general physiological conditions. Although lacking specific structures, disordered regions play crucial functional roles in many biological processes, such as transcriptional regulation, translation, cellular signal transduction, and others; they are also shown to be prevalent in various human diseases including cancer, cardiovascular disease, and genetic diseases, to name a few. Due to their functional importance, it is with great urgency that we search for a way to reliably identify disordered regions. Experimental techniques for disorder detection are costly and time-consuming, and for some proteins currently impossible. Thus, many computational methods have been developed for predicting disordered regions. Combining different methods often results in improved prediction accuracy since different methods have information arising from different sequence features and different training data. Here, we selected six existing disorder prediction methods and built a meta-predictor, SPINE-DM. Preliminary results suggest that SPINE-DM provides accurate predictions of disordered regions.

Methods

A set of 1229 protein chains, named DP1229, was selected for training and testing SPINE-DM. It was built based on a set of X-ray-determined structures in the PDB having residues without atomic coordinates, as well as the set of fully disordered proteins in the Disprot[1] database. This set is further culled to the DP1229 dataset by restricting the pairwise sequence identity to below 25%.

We used a two-hidden-layer neural network with a filter predictor for smoothing the predictions, a hyperbolic activation function and guided learning technique developed for Real-SPINE 3.0[2]. Considering that the primary sequences in long and short disordered regions are dissimilar, we sorted residues into 3 classes: ordered residues, residues in short disordered regions (<= 30 residues) and residues in long disordered regions (>30 residues). Later, predictions for the latter two classes were combined to yield the final disorder predictions. To reduce random prediction errors caused by the randomly selected initial weights, we trained five independent predictors and the final prediction is based on their consensus.

We considered several well-known disorder predictors. The main prerequisite for the methods considered was that they must offer a standalone implementation that can be incorporated into local predictive pipelines. Six predictors were considered in total, including VSL2[3], Disopred2[4], Dispro1.0[5], IUPred[6] (two versions: IUPredS for short disorder and IUPredL for long disorder) and SPINE-D (our own predictor). All predictors provided both binary (order/disorder) and confidence values (probability of a prediction to be correct). The confidence values from the six predictors were used as inputs for the neural networks.

We applied a cost matrix to accommodate for the imbalanced populations of disordered and ordered residues, i.e., higher weights were given to the minority (disordered) residues. More specifically, we made duplicates for disordered samples during the training process.

The window size and parameters were optimized for the highest AUC (area under the ROC curve) value. Predicted probabilities for order/disorder residues were further optimized by selecting a threshold that led to the highest S_w score[7] in binary predictions.

Results

A ten-fold cross validation test was performed on the DP1229 dataset. Our method achieved an AUC of 0.865 and a S_w of 0.599, which outperforms each of the six used methods.

Our final predictor, SPINE-DM, was trained on the entire DP1229 dataset. We also tested the SPINE-DM server on the CASP8 dataset. The AUC was 0.905, and S_w equaled 0.690.

Availability

SPINE-DM is available at http://sparks.informatics.iupui.edu/SPINE-DM/index.html

1. Sickmeier M., Hamilton J.A., LeGall T., Vacic V., Cortese M.S., Tantos A., Szabo B., Tompa P., Chen J., Uversky V.N., Obradovic Z., Dunker AK. (2007). DisProt: the Database of Disordered Proteins. *Nucleic Acids Res.* **35** (Database issue): D786-93.
2. Faraggi E., Xue B., Zhou Y. (2009) Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two-layer neural network. *Proteins* **74**: 847–856.
3. Peng K., Radivojac P., Vucetic S., Dunker A.K., Obradovic Z. (2006) Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics*. **7**:208.
4. Ward J.J., Sodhi J.S., McGuffin L.J., Buxton B.F., Jones D.T. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* **337**(3):635-45.
5. Cheng J., Sweredoski M., Baldi P. (2005) Accurate Prediction of Protein Disordered Regions by Mining Protein Structure Data. *Data Mining and knowledge Discovery*. **11**(3):213-222.
6. Dosztányi Z., Csizmok V., Tompa P., Simon I. (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*. **21**(16):3433-4.
7. Jin Y., Dunbrack R.L. Jr. (2005) Assessment of disorder predictions in CASP6. *Proteins*. **61** (Suppl 7) :167-175

AUTHOR INDEX

F

G

H

I

J

K

L

M

T

U

V

W

X

Y

Z