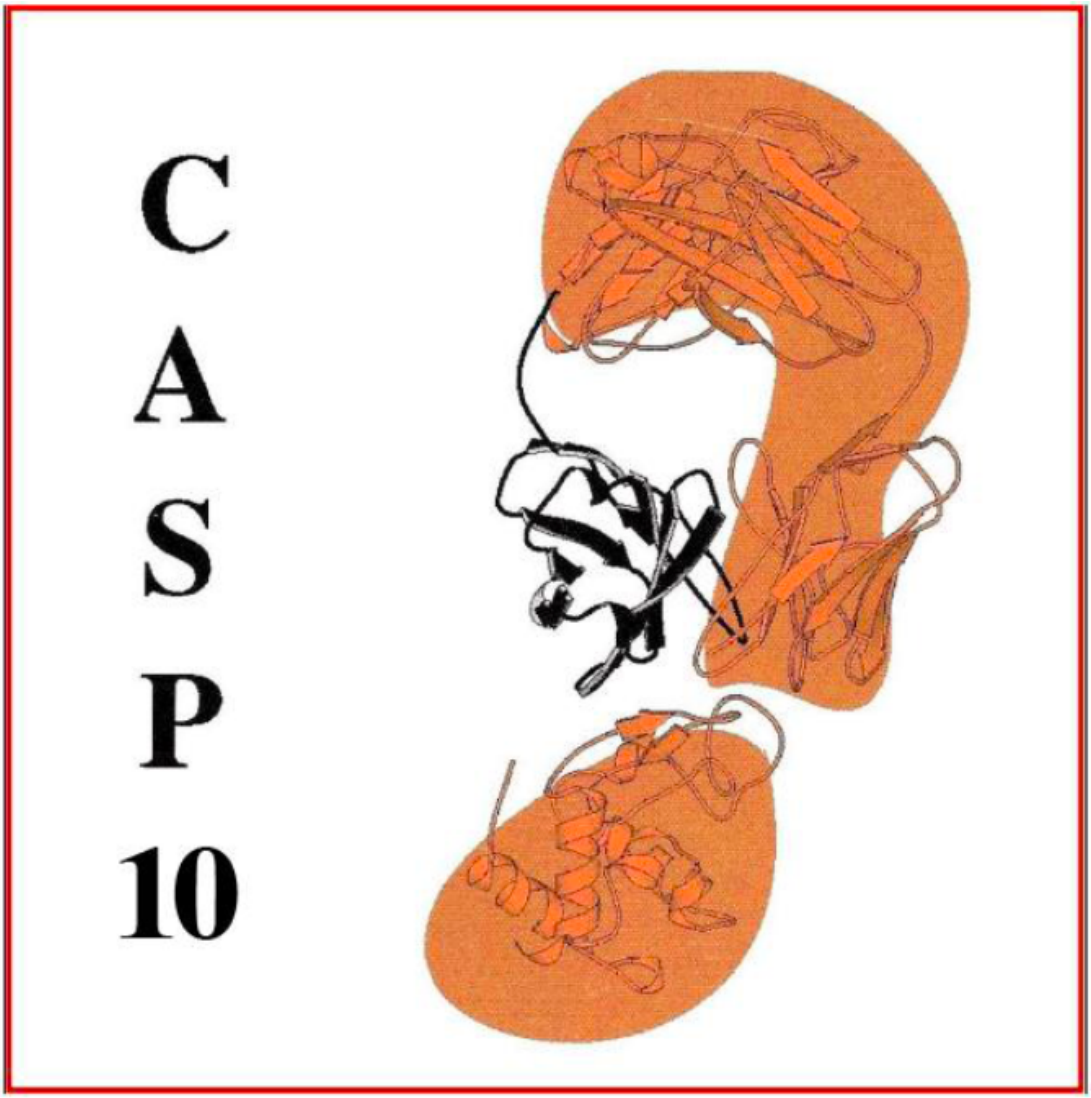
The EMBO Conference: Critical assessment for protein structure prediction (CASP10)



Gaeta, Italy
December 9-12, 2012

TABLE OF CONTENT

AIdisorder – protein disorder prediction using machine learning and trivial features

L.P. Kozlowski[1] and J.M. Bujnicki[1]

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

AIdisorder is an experiment to check whether adding trivial features, easy to extract or predict form amino acid sequence (e.g. secondary structure, homology information, etc.) to state-of-the-art disorder predictor can improve disorder prediction accuracy.

Methods

AIdisorder is a combination of two methods. The first is sDisPred protocol (described in more details in other abstract in CASP10) and the second is GSmetaDisorderMD[1] server. In brief, sDisPred is a simple predictor of disorder based on easy to predict features like the existence of similar structures in PDB[2] and DISPROT[3] databases, secondary structure, some statistical functions and finaly the consensus from already available disorder predictors. On the other hand, GSmetaDisorder is a sophisticated machine learning metapredictor which uses information from over 20 primary disorder predictores, secondary structure and fold recognition programs. The two parts were combined using genetic algorithm.

Results

Our internal benchmark made by using CASP8 and CASP9 targets shows that the combination of sDisPred and GSmetaDisorder gives an added value to the disorder prediction.

Availability

The method will be publicly available in the form of a web service if it proves to be valuable in terms of disorder prediction in current CASP.

1. Kozlowski,L.P. & Bujnicki,J.M. (2012). MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins. BMC Bioinformatics 13, 111.
2. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. & Bourne,P.E. (2000). The Protein Data Bank. Nucleic Acids Res 28, 235-42.
3. Vucetic,S., Obradovic,Z., Vacic,V., Radivojac,P., Peng,K., Iakoucheva,L.M., Cortese,M.S., Lawson,J.D., Brown,C.J., Sikes,J.G., Newton,C.D. & Dunker,A.K. (2004). DisProt: a database of protein disorder. Bioinformatics.

Prediction of Disordered Regions of Proteins Using Code of Alpha-Helical Structure

B.V. Shestopalov
Institute of Cytology of Russian Academy of Sciences
bvshest@mail.cytspb.rssi.ru

The method used for prediction of disordered regions in proteins is developed using results of researches aimed to discover a code of protein structure[1-5]. It is postulated that such code is algorithmic one.

Methods

It is proposed that disordered regions in proteins are regions where alpha-helices can not be formed.

A code of the alpha-helical structure is developed and alpha-helical domains are predicted using this code. All the regions outside these domains are disordered ones.

All the predictions were done by the same method but little correction, resulted in more length of some disordered regions, was made for predictions submitted from 2012-06-27 (T0704, T0705, T0707, T0713, T0717, T0719, T0721-T0758).

Mixed, manual and computer, calculations were done but in manual case strict programmable rules were used.

Availability

Detailed description of the method, illustrated by results of its application in the CASP10 experiment will be published.

1. Shestopalov,B.V. (1990). Predictioon of protein secondary structure by doublet code method. *Mol. Biol.* **24**, 1117-1125.
2. Shestopalov,B.V. (2003). Amino acid code of protein secondary structure. *Tsitologiya*. **45**, 702-706.
3. Shestopalov,B.V. (2003). Statistical model of amino acid code of protein secondary structure. *Tsitologiya*. **45**, 707-713.
4. Shestopalov,B.V. (2007). The code-based physics of formation of α-helices and β-hairpins in water-soluble proteins. *Doklady Biochemistry and Biophysics*. **416**, 245-247.
5. Shestopalov,B.V. (2007). Simulation of formation of α-helices and β-hairpins in water-soluble proteins by the code-based physics. *Cell and Tissue Biology*. ***1***, 420-426.

Multiplayer Online Game-Based Homology and Ab-Initio Modeling

F. Khatib[1], J. Flatten[1], S. Cooper[1], T. Husain[1], K. Xu[1], Z. Popović[1], D. Baker[1] and Foldit Anthropic Dreams Group[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. CASP10 targets shorter than 170 residues were given to Foldit players as puzzles to solve. Foldit allows players to form groups for cooperative gameplay; in this case predictions were selected from and by members of the Foldit group Anthropic Dreams.

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 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 CASP10 puzzle was typically accessible to Foldit players for 8-9 days.

For CASP10 targets shorter than 170 residues in the "All Groups" category, two different Foldit puzzles were given to the players. One puzzle started from an extended chain, with alignments to known templates taken from the RAPTOR[3], SPARKS[4], and HHsearch[5] servers provided. 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 in-game tools listed above. For the second puzzle, models were constructed using the QUARK[6] and Zhang-Server[7] predictions. These server models were initially minimized using Rosetta and then given as starting points for the Foldit players to refine. This same protocol was used for CASP10 targets in the "Refinement" category, where server models were first minimized with Rosetta before being given to the Foldit players. Foldit players were provided with secondary structure predictions, generated by the SAM-T08 server[8], in the form of a sequence logo for all CASP10 puzzles.

Quality and ranking of individual models was determined initially by the Rosetta full-atom energy. Submissions were then selected from high- and medium-ranking Anthropic Dreams predictions based on the fit between actual difference (RMSD or GDT_TS) of the prediction from the starting model and expected difference of a good solution from the starting model. Additional selection criteria included conformational diversity among submissions and diversity of players represented.

Availability

Foldit is available through the Rosetta Commons at http://tinyurl.com/academic-foldit

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. *Methods in Enzymology* **487**, 545-74.
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. Peng,J. & Xu,J. (2009) Boosting Protein Threading Accuracy. *Research in Computational Molecular Biology (RECOMB)* **5541**, 31-45.
4. Yang,Y., Faraggi,E., Zhao, H. & Zhou,Y. (2011) Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of the query and corresponding native properties of templates. *Bioinformatics* **27**, 2076-2082.
5. Söding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**(7), 951-60.
6. Xu,D. & Zhang,Y. (2012) Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins* **80**, 1715-35.
7. Zhang,Y. (2008) I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* **9**, 40.
8. Karplus,K. (2009) SAM-T08: HMM-based Protein Structure Prediction. *Nucleic Acids Research* **37**(2), W492-7.

Protein structure modeling guided by homology and hydrophobic residue interactions

Matsuyuki Shirota

Tohoku University
mshirota@hgc.jp

AOBA-server is an automated server for protein structure modeling. For targets with significant homologues, the models were made using comparative modeling. In cases with no homologous templates the models were made to form hydrophobic contacts between local segments.

Methods

We downloaded target amino acid sequences from the CASP website. For each of the sequences, homology search was performed in PDB using HH search program[1]. If significant hits for the entire sequence were found, 3D models were made using MODELLER program[2] based on the alignments and the top five models determined by the alignment scores were submitted. If the sequence included some regions which did not have homologous template structures, we first modeled the aligned segments based on the template structure and then modeled the unaligned regions. These regions were assumed to take secondary structures as predicted by PSIPRED[3], and to form hydrophobic contacts with the aligned domain with its exposed hydrophobic residues. About 100~1,000 models were generated by MODELLER under the constraints of secondary structures and distance restraints between hydrophobic residues. These models were evaluated by a scoring function, which includes median TM-score[4] to all the other models and Verify3D[5] score and 3D-1D stability score[6], and the top five structures were submitted. If there were no significant hit for the target sequence, models were generated using weakly homologous local templates found by HH search. These local templates, together with secondary structure predictions, were used as restraints to model the protein structure. We generated about 1,000 structures for each target and they were evaluated as the same method as described above.

1. Soding,J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**(7):951-960.
2. Sali,A. & Blundel,T.L. (1993). Comparative protein modeling by satisfaction of spatial restraints. J. Mol. Biol. 234, 779-815
3. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* ***292***, 195-202.
4. Zhang,Y., &Skolnick,J. (2004). Scoring function for automated assessment of protein structure template quality. Proteins. 57, 702-710
5. Bowie,J.U., Luthy,R., & Eisenberg,D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. Science 253, 164-170.
6. 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.

Meta-methods for model quality assessment

M. J. Skwark[1,2]

1 – Department of Biochemistry and Biophysics, Stockholm University,
2 – Science for Life Laboratory, Stockholm
marcin.skwark@scilifelab.se

Ariadne and ConQ are model quality assessment methods, that combine multiple MQAP approaches in order to avoid method biases and consequently provide more impartial ranking of predicted models.

Methods: Ariadne

Ariadne as a model quality assessment method aims to identify most accurate models in the model ensemble by performing structural comparison of all the models to a putatively selected set of models of good quality. The comparison set is chosen by scoring the model ensemble in question by multiple quality assessment approaches: Pcons[1] ProQ2[2], PconsD and dDFIRE[3]. For each of the compound methods 5 highest ranking models are chosen to be included in the *comparison set*. Afterwards all the models in the ensemble are compared by means of TM-score to the models in *comparison set*.
The final score is a mean of the superposition scores, discarding 20% of outlier results.

Methods: ConQ

ConQ as a model quality assessment method aims to incorporate sequence-based contact prediction into the model scoring framework. Using the same MQAP methods as Ariadne (Pcons[1], ProQ2[2], PconsD and dDFIRE[3,4]) models are assigned ordinal rank-based scores (for n models the top ranked model gets score n, the next one score n-1 etc.) for each MQAP method.
Additionally models are ranked based on the agreement of detected inter-residue contacts in the model with the predicted contact maps. Contact maps are predicted using DCA[5] and Psicov[6] methods, based on JackHMMer[7] alignments to a UniRef100 database.
Final score is then normalized for the highest ranked model to get score 1 and lowest ranked – score 0.
Models submitted by **Ariadne** and **ConQ** as a manual structure prediction methods are the ones which earn the highest score for respective MQAP methods

1. Larsson P., Skwark MJ, Wallner B and Elofssson A (2009) Assessment of global and local model quality in CASP8 using Pcons and ProQ2 Proteins 77 (S9): 167-172
2. Ray A., Lindahl E. and Wallner B. (2012) Improved model quality assessment using ProQ2 BMC Bioinformatics 13, 224-
3. Yang Y. and Zhou Y. (2008) Specific interactions for ab initio folding of protein terminal regions with secondary structures Proteins (72) 798-803
4. Yang Y. and 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 (17) 1212-1219
5. Morcos F. et al. (2011) Direct-coupling analysis of residue coevolution captures native

contacts across many protein families PNAS (108) E1293-E1301
6. Jones DT, Buchan DWA, Cozzetto D and Pontil M. (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments Bioinformatics (28) 184-190

@TOME-2: a pipeline for comparative modeling of protein-ligand complexes

Jean-Luc Pons, Jérome Gracy, Gilles Labesse

Centre de Biochimie Structurale, MONTPELLIER France

@TOME 2.1(1) is a web pipeline dedicated to protein structure modeling and small ligand docking based on comparative analyses. @TOME 2.1 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 connected the comparative docking of small ligands which 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 CASP10 were automatically treated as follows:

The best structural alignments (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 software (6).

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

In a second step, 24 multi-template models were computed by MODELLER 9.0 (12). For each model to construct, 2, 3 and 4 templates have been selected according the best scores from @TOME-2, Verify3D, TITO and Qmean. Each model is calculated with and without conserved 3D restraints calculated by the FCT tool from the PAT software (13). The restraints correspond to the most frequent atomic contacts observed in the superimposed structures. Among the 24 obtained models, the 5 best QMean scores have been proposed to CASP10.

Moreover, comparative docking has been used for the the automated detection of active sites.

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

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. Gracy,J. Chiche,L. (2005). « PAT: a protein analysis toolkit for integrated biocomputing on the web », Nucleic Acids Res 33 (Web Server issu) W65-71

Modeling of Protein Structures Using Rosetta in CASP 10

T. Brunette*, R. Wang*, D.E. Kim*, F. DiMaio, S. Ovchinnikov, K. Jung, H. Kamichetty, Y. Song, F. Khatib, C. Miles, J. Thompson, D. Baker

University of Washington, Seattle, WA
dbaker@uw.edu

In CASP 10, the BAKER human group evaluated new protocols in the template-based modeling, refinement and contact assisted categories. For comparative modeling we used a new template hybridization method described in the BAKER-ROSETTASERVER abstract, but added alignments from servers and domain parsing based on human inspection. For assisted structure prediction targets, we identified sub-alignments, fragments, and *de novo* models that satisfied the provided contacts and used them as input for our hybridization protocol. For refinement we used a new search procedure that found much lower full-atom energy structures by quickly building loops and batch minimizing the energy of structures.

Methods

Template Identification and Domain Parsing

For Baker human group, templates were identified both by the a suite of locally-installed threading programs (HHSearch[2], Sparks-X[3], RaptorX[4]) and by searching through the protein data bank for proteins that best match servers models using TMalign[5]. Using these alignments each target was parsed into domains. When no obvious template was available domain parsing was done automatically using GINZU[6]. For some cases (for example, T0651 and T0674), domain predictions were made by human inspection of alignments and results from the NCBI conserved domain database.

Structure Assembly

When templates were available, our new template hybridization protocol was used for modeling. For multi-domain targets, individual domains were folded separately and then assembled by energy-guided rigid docking followed by loop closure within the hybridization protocol.

Free Modeling and CASP Roll

In cases with no clear similarity to known structures we used the standard Rosetta *de novo* fragment-assembly approach. Fragment selection was improved through the incorporation of predicted torsion angles and solvent accessibility using Spine X[7], and fragment based structure profiles[8]. For each target both the original sequence and several homologous sequences were modeled generating approximately 100k-300k decoys. The lowest energy 4000 decoys were structurally clustered, and the lowest scoring full-atom decoy from each cluster was returned as the final prediction.

Contact Assisted

For assisted structure prediction targets, we identified putative templates that best satisfied the provided contacts for inputs to the hybridization protocol from (a) sub-alignments from the top 1000 Sparks alignments, (b) PDB fragments from fragment libraries used for *de novo* modeling, and (c) models generated with the *de novo* structure prediction protocol using the provided

contacts as constraints during sampling. Final models from the hybridization and *de novo* structure prediction protocols were selected based on satisfied contacts and Rosetta energy.

Refinement
The refinement procedure used a new sampling strategy called loophash[9] to produce decoys with very low full-atom energy. Models were selected based on the criteria that they were low energy, structurally diverse and fell close but above the given starting GDT. Initial results suggest our refinement method produced structures closer to native than our other sampling approaches, but the selection procedure chose decoys too far from native.

Results
Most of the targets where our human group significantly outperformed our server are in the hard regime, for example T0651 and T0724. In the hard regime accuracy is largely determined by the quality of templates identified, indicating the human group was using better templates. Modeling failures resulted from inclusion of inaccurate templates, poor model selection and incomplete sampling of the rugged Rosetta full-atom energy landscape. For contact guided targets, our sampling strategies produced models that satisfied all provided contacts and generally improved over our server and human predictions.

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

Acknowledgements
We thank Jinbo Xu, Johannes Söding, Yaoqi Zhou for making their excellent software available.

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. *Methods in enzymology* 487, 545-74.
2. Söding,J. (2005) *Protein homology detection by HMM-HMM comparison. Bioinformatics.* 21(7):951-60
3. Yang,Y., Faraggi,E., Zhao, H. & Zhou,Y. (2011) Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of the query and corresponding native properties of templates. *Bioinformatics* 27, 2076-2082.
4. Peng,J. & Xu,J. (2009) Boosting Protein Threading Accuracy. *Research in Computational Molecular Biology (RECOMB)*, 5541, 31-45.
5. Zhang,Y & Skolnick,J. (2005) TM-align: A protein structure alignment algorithm based on TM-score, Nucleic Acids Research 33:2302-2309)
6. 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.
7. Faraggi E., Zhang T., Yang Y. Kurgan L., Zhou Y. (2012) SPINE X: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles.33(3)259-67

8. Zhou H, Zhou Y. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins 2005;58:321–328.
9. Tyka, M., Jung K. & Baker D. (2012).Efficient Sampling of Protein Conformational Space using Fast Loop Building and Batch Minimization Highly Parallel Computers. Journal of Computational Chemistry. July

Robetta 2.0: Full-chain protein structure prediction server for CASP 10

C.B. Miles[1], D.E. Kim[2], T. Brunette[2], R.Wang[2], Y.F. Song[2], F.P. DiMaio[2], D. Baker[2]

[1]*Department of Computer Science and Engineering*

[2]*Department of Biochemistry, University of Washington, Seattle, WA 98195*

dabaker@uw.edu

In CASP10, we evaluated a new protocol for comparative modeling that assembles models through recombination of structural elements present in multiple templates. This approach improves upon the LoopRelax method used in previous CASP experiments by (1) leveraging orthogonal information present in multiple templates within a single trajectory and (2) replacing cyclic coordinate descent loop closure with a combination of fragment insertion and mixed torsion-Cartesian space minimization. Both operations are carried out using a centroid level representation of the polypeptide chain, with the best-scoring models relaxed in Rosetta's full-atom forcefield. Benchmarks performed prior to CASP10 suggested that this new method of model building produced more accurate results for targets in the Easy and Medium difficulty regimes. In the Hard regime, results are less sensitive to model building, since accuracy is largely determined by the quality of templates identified.

Methods

Template Identification. Impressed by the progress in template identification demonstrated in CASP9, we employed a suite of locally-installed threading programs (HHSearch[1], Sparks-X[2], RaptorX[3]) to generate alignments. From the sets of alignments produced by these methods, we identified domain boundaries and assessed target difficulty based on the degree of structural consensus among each methods' top predictions. The threading models were then clustered to identify distinct topologies, which were ranked based on the likelihood of the constituent alignments. Spatial restraints were generated separately for each cluster[4]. Symmetry information was inferred from the top predictions from each method.

Structure Assembly. The template hybridization protocol operates in three distinct phases. Beginning from a randomly selected template in the alignment cluster, the first phase samples alternative global topologies by inserting continuous chunks excised from the partial threading models. Unaligned loop regions are rebuilt de novo. Both operations are performed under Rosetta's low-resolution energy function using novel, broken-chain kinematics, which limit the extent of conformational propagation. In the second phase of the protocol, chunk insertions alternate with Cartesian-space minimization to refine local geometry, particularly backbone hydrogen bonding networks. In the final phase, the predicted structure is refined using Rosetta's full-atom energy function. For difficult targets, models generated by the template hybridization protocol are supplemented with models generated using Rosetta's ab initio protocol[5]. The number of models generated for each topological alignment cluster increases with predicted target difficulty. All structures were generated on Rosetta@Home.

Model Selection. Robetta employs a hierarchical model selection procedure: final models are chosen by combining individual selections performed on each topologically distinct alignment cluster. After performing an initial energy cut, selection proceeds by repeatedly selecting the model that minimizes the Rosetta full-atom energy and distance to the remaining models in the cluster. After that model is selected, the most structurally similar 10% of the remaining models are also removed from consideration. The order of final models is determined by the summed likelihood of the alignment cluster to which they belonged. The decision to submit symmetric models is determined based on the $\Delta\Delta G$ of the complex normalized by the number of subunits.

Results

Using the template hybridization protocol, we produced models for several targets that were clear improvements over the best starting template identified, notably T0662, T0667, and T0714. 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.

Availability

Robetta is available for non-commercial use at http://robetta.bakerlab.org. Source code for the template hybridization protocol is included in the upcoming Rosetta 3.5 release, which can be downloaded from http://www.rosettacommons.org.

1. Söding,J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21 (7)**, 951-960.
2. Yang,Y., Faraggi,E., Zhao,H. & Zhou,Y. (2011). Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of the query and corresponding native properties of templates. *Bioinformatics* **27 (15)**, 2076-2082.
3. Peng,J. & Xu,J. (2011). Raptorx: Exploiting structure information for protein alignment by statistical inference. *Proteins* **79,** 161-171.
4. Thompson,J. & Baker,D. (2011). Incorporation of evolutionary information into Rosetta comparative modeling. *Proteins* **79 (8)**, 2380-2388.
5. 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. *Methods in Enzymology* **487**, 545-574.

SwarmLoop: a program to dock fragments and domains

R.A.G. Chaleil and P.A. Bates

Cancer Research UK London Research Institute
paul.bates@cancer.org.uk

Our objective is to find the global energy minimum for native protein structures by blending diverse sets of models, created by several different means. We have developed a number of refinement operators (the move-set) to search restricted region of conformational space. These operators are embedded within a genetic algorithm (GA) that reshuffles and repacks structural components[1-2]. The most recent addition to our methodology is to add a torsion angle optimisation method to pack protein fragments and domains.

Methods

For CASP10 we entered a fully automated server, 3D-JIGSAW_V5-0, employs our GA approach, and our new refinement algorithm call, SwarmLoop, in the manual intervention section. Potential templates and fragments are first identified using the HHpred package[3].

All templates are modelled to the target sequence using the side-chain replacement program SCWRL[4]. Insertions and deletions are modelled by our in-house loop modelling and closure method[2] a modified version of the cyclic coordinate descent algorithm[5].

The initial population of models is ranked with our coarse-grain energy function[2] before being fed into rounds of GA optimization. Rounds of GA optimization employ the principles of crossover, mutation and model selection as previously described[1-2].

For manual and refinement predictions we employed our new algorithm, SwarmLoop. The basic ideas for this algorithm have been translated from our approach for docking proteins[6].

SwarmLoop, is a memetic optimisation algorithm in which parameters are simultaneously optimised using the Particle Swarm optimisation (PSO) metaheuristic[7].

In SwarmLoop, the PSO optimises:

1. The phi and chi torsion angles of selected residues (loop residues linking fragments or domains)
2. The chi1 and chi2 side-chain torsion angles of residues on the surface of the fragments.
3. The coefficients in the linear combination of the five lowest frequency normal modes of each fragment; the normal modes are Calculated prior to the optimisation using ElNemo software[8].

Backbone torsion angles are selected from a statistical distribution obtained from the structural database. The distribution was calculated by clustering a non-redundant set of structures. The probability of emission of torsion angles in the PSO matches that of the Ramachandran plot.

The energy function is an optimised reimplementation of the Dfire[9] pair potential. Missing loops between fragments are built *ab-initio* with random torsion angles and then optimised with the SwarmLoop algorithm.

To check to see if our repacked structures were likely to be functional, we performed

numerous literature searches on closely homologous proteins.

Availability

Http://bmm.cancerresearchuk.org/~populus/

1. Offman M.N., Fitzjohn P.W. & Bates P.A. (2006) Developing a move-set for protein model refinement. Bioinformatics. 22, 1838-1845.
2. Offman M.N., Tournier, A.L. & Bates, P.A. (2008) Alternating evolutionary pressure in a genetic algorithm facilitates protein model selection. BMC Struct. Biol. 8:34.
3. Soding, J., Biegert, A. & Lupas A.N. (2005) The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res., 33, W244-W248.
4. Canutescu, A.A., Shelenkov, A.A. & Dunbrack R.L.Jr. (2003) A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci., 12 , 2001-2014.
5. Canutescu, A.A. & Dunbrack R.L.Jr. (2003) Cyclic coordinate descent: a robotics algorithm for protein loop closure. Protein Sci.,12, 963-972.
6. Moal I.H. & Bates P.A. (2010). SwarmDock and the use of normal modes in flexible protein-protein docking. Int. J. Mol. Sci.,11, 3623-3648.
7. Kennedy, J. & Eberhart, R.C. Partical Swarm Optimization. In Proceedings of the IEEE International Conference on Neural, Peth, Australia, 1995; 4,1942-1948.
8. Suhre, K. & Sanejouand, Y.H., ElNemo: a normal mode web-server for protein movement analysis and the generation of templates for molecular replacement. Nucleic Acids Research, 32, W610-W614, 2004.
9. Y. Yang and Y. Zhou, ``Specific interactions for *ab initio* folding of protein terminal regions with secondary structures.", *Proteins* **72**, 793-803 (2008)

Bhageerath-H a homology/ ab-initio method for predicting tertiary structures of soluble monomeric proteins

Priyanka Dhingra[1,3], Avinash Mishra[2], Satyanarayan Rao[3] and B. Jayaram[1,2,3]

[1] – Department of Chemistry, [2] –Kusuma School of Biological Science, [3] –Supercomputing Facility for Bioinformatics and Computational Biology, Indian Institute of Technology Delhi, India

bjayaram@chemistry.iitd.ac.in, priyanka@scfbio-iitd.res.in

Protein folding, considered to be a grand challenge problem in modern science and a holy grail of molecular biology, remains intractable even after six decades since the report of the first crystal structure. The advent of human genome sequencing project has led to an explosion in the number of protein sequences in databanks. Despite the availability of over 80,000 X-ray and NMR structures in the RCSB protein data bank, there is a diverging gap between available sequences and structures, which calls for an immediate *in silico* solution. Computational methods such as homology modeling which rely on extracting information from the known structures in PDB have proved to be successful in predicting tertiary structures of sequences which share high sequence similarities. With dwindling similarities of query sequence with databases, newer *ab initio* / homology hybrid approaches are being explored to solve the structure prediction problem. In an effort to drive the prediction accuracies beyond the current limit, we have developed *Bhageerath-H* (www.scfbio-iitd.res.in/bhageerath/bhageerath_h.jsp) a homology/ *ab initio* hybrid software for predicting tertiary structures of soluble monomeric proteins.

Methods

Bhageerath-H is a homology/*ab initio* hybrid method for protein tertiary structure prediction. It takes input amino acid sequence and provides as an output five candidate structures for the native. The software initially predicts secondary structure of the input query sequence and searches the PDB, Pfam and SCOP databases for sequence and family based homologs. It then uses softwares such as pGenthreader[1], HHSearch[2] and ffas[3] for finding templates and generates target-template alignments. Each of the selected templates and target-template alignments are used for generating a library of 3D protein structural models of varying length, sizes and folds. Missing residue stretches are searched in the modeled patches and generated using Bhageerath[4-7] *ab initio* 3D modeling software and the top five Bhageerath *ab initio* energy ranked structures are incorporated in the growing library. The incomplete protein patches in the protein model library are patched in all possible combinations with the remaining protein patches to put forth complete models, which then undergo few cycles of energy minimization. Using pcSM[8] a physico-chemical scoring metric, which comprises parameters such as intra molecular energy, accessible surface area, euclidean distance and secondary structure propensity for detecting native and non-native like structures in the decoy pool, top 100 complete structures are selected. These selected structures are later optimized using Bhageerath *ab initio* loop modeling and the pcSM selected top five structures are finally processed via Monte Carlo based side chain modeling and short MD simulations.

Results

The two main issues involved in the success of a protein tertiary structure prediction algorithm are sampling and scoring. Are we sampling the conformational space enough to generate native-like structures? How best can we distinguish a native-like structure from a non-native-like structure? To answer these questions we have tested the ability of Bhageerath- Strgen[9] method for generating native like decoys on the benchmark CASP9 dataset of 116 targets. In 93% of the cases, a structure with TM-score ≥ 0.5 is generated in the pool of decoys. pcSM scoring method was tested on a dataset of 415 systems and 142698 decoys and is able to detect native or native like structures in the top five with 93% accuracy from an ensemble of candidate structures.

Bhageerath-H has been fielded in CASP10 in the server category. Native structures of 38 targets have been released so far. *Bhageerath-H* is able to generate a native like structure to within 3 Å in 20 cases, the best server count being 23. Several improvements are being considered for implementation in the near future to push the accuracies even higher.

Availability

The protocol has been web enabled and is freely accessible at www.scfbio-iitd.res.in/bhageerath/bhageerath_h.jsp.

1. Lobley,A., Sadowski,M.I. & Jones,D.T. (2009). pGenTHREADER and pDomTHREADER: New Methods For Improved Protein Fold Recognition and Superfamily Discrimination. *Bioinformatics* **25,** 1761-1767.
2. Soding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics* 21, 951-960.
3. Jaroszewski,L., Rychlewski,L., Li,Z., Li,W. & Godzik,A. (2005) FFAS03: a server for profile--profile sequence alignments. *Nucleic Acids Res*. **33**, W284-288.
4. 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.
5. Jayaram,B., Bhushan,K., Shenoy,S.R., Narang,P., Bose,S., Agrawal,P., Sahu,D. & Pandey,V. (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.
6. Narang,P., Bhushan,K., Bose,S. & Jayaram,B. (2006) Protein structure evaluation using all-atom energy based empirical scoring function. *J Biomol Struct Dyn*. **45A**, 1834-1837.
7. Jayaram,B., Dhingra,P., Lakhani,B. & Shekhar S. (2012) Bhageerath - Targeting the Near Impossible: Pushing the Frontiers of Atomic Models for Protein Tertiary Structure Prediction. *Journal of Chemical Sciences* **124**, 83-91.
8. Mishra,A., Rao,S., Mittal,A. & Jayaram,B. Capturing Native/Native like protein structures with a physico chemical metric-(pcSM) (manuscript submitted)
9. Dhingra,P., Lakhani,B. & Jayaram,B. Generating native-like structures of soluble monomeric proteins via Bhageerath-H a homology/ *ab initio* hybrid method. (manuscript submitted)

Tertiary Structure Prediction by Combination of Fold Recognition and Fragment Assembly and Semi-automated Quaternary Structure Prediction from Monomer Structure

S. Nakamura

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

Bilab-ENABLE is a fully-automated prediction server and Bilab is a human prediction group which generated a quaternary structure model from a server model.

Methods

Overview of the procedure of our ENABLE server is as follows: 1) Template search by PDB-BLAST and HHpred[1] combined with T-COFFEE[2]. 2) Generation of the 1st set of models by using MODELLER[3] from a variety of template-alignment combinations. 3) Fragment assembly called IDDD/ABLE[4] developed in our laboratory using fragments from models of the 1st set and structures in PDB. 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. Generated models were added to the 1st set. 4) Top 500 models were selected according to Verify3D[5] 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. 5) Predicted quaternary structures were generated using the similar procedure as group Bilab described below using the first monomer model.
Group Bilab picked up the most probable model from server models and generate quaternary structure model by a template-based method. In this method, templates for quaternary prediction based on HHpred were selected according to score calculated by alignments. Predicted monomer structure was then superimposed to the template quaternary structure. Finally, generated quaternary models were evaluated and best scored quaternary models were submitted.

1. Soding,J. (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics 21, 951-960.
2. 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.
3. Sali,A. & Blundell,T.L. (1993) Comparative protein modeling by satisfaction of spatial restraints. J. Mol. Biol. 234, 779-815.
4. 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.
5. Luthy,R., Bowie,J.U. & Eisenberg,D. (1992) Assessment of protein models with three-dimensional profiles. Nature 356, 83-85.

An Integrative Method for Prediction of Ligand Binding Residues

Yi Xiong[1,2§], Ishita Khan[2,1§] and Daisuke Kihara [1,2*]

[1] *Department of Biological Sciences;*

[2] *Department of Computer Science, Purdue University, USA*

[§]These authors contribute equally to the work

[*]Corresponding author: dkihara@purdue.edu

In CASP10, we designed a new integrative method based on our in-house software and two publicly available web servers, all of which were developed previously in our group. They are structure based ligand binding residue prediction tool SiteHunter and sequence based function prediction methods Protein Function Prediction (PFP)[1] and the Extended Similarity Group (ESG)[2].

Methods

We developed an integrative method to predict ligand binding sites. SiteHunter is constructed based on the observation that homologous proteins tend to bind ligands at similar binding residues. Firstly, a query (also called target here) protein sequence was submitted to the Pfam database to identify homologous proteins in the same Pfam family. Homologous proteins were checked whether they have known structures in PDB with bound ligands. Then, the homologous proteins with bound ligands were aligned with the query protein using MUSCLE. Finally, the ligand binding locations on homologous structures were transferred to the corresponding location in the query structure by the alignment result.

In order to identify ligand binding residues based on sequence similarity, we used our in-house tools PFP and ESG, both of which are sequence similarity based automated function prediction (AFP) methods. The strength of PFP is its coverage of a large number of sequences, by including weakly similar sequences into consideration. On the other hand, ESG assigns higher weight on the consensus sequences that have strong similarity with the query protein among all the sequences it encounters along multiple iterations. Based on the PFP and ESG predictions of a target binding to a significant ligand, we found the template sequences (ranked according to the sequence similarity with the target protein) that have been retrieved by the PFP and ESG to make the prediction in question. Then we found the ligand binding residues for the top template sequences from Uniprot and found the corresponding binding residues of the target from multiple sequence alignment using MUSCLE. For some top hit sequences that did not have the binding residues listed in Uniprot, we used some ligand specific servers such as PredZinc, ATPsite, NADbinder to identify the ligand binding residues.

On the query sequence, the binding sites predicted from the individual homologous proteins were integrated by majority voting.

Results

Our integrated method made predictions of ligand binding sites on a total of 110 targets in CASP10, which are categorized as 'All groups' or 'Server only'. We predicted 83 (75.5%) of these 110 targets with bound ligand.

Availability

http://kiharalab.org/pfp.php
http://kiharalab.org/esg.php

Acknowledgements

This work has been supported by grants from the National Institutes of Health (R01GM075004, R01GM097528), National Science Foundation (EF0850009, IIS0915801, DMS0800568), and National Research Foundation of Korea Grant funded by the Korean Government (NRF-2011-220-C00004).

1. Hawkins,T., Luban,S. & Kihara,D. (2006). Enhanced automated function prediction using distantly related sequences and contextual association by PFP. *Protein Sci*. **15**, 1550-1556.
2. Chitale,M., Hawkins,T., Park,C. & Kihara,D. (2009). ESG: extended similarity group method for automated protein function prediction. *Bioinformatics*. **25**, 1739-1745.

biomine {biomine_DR_mixed, biomine_DR_pdb, biomine_DR_pdb_c}

Prediction of Disordered Regions by Multilayered Information Fusion

M.J. Mizianty[1], T. Zhang[2], Z-L. Peng[1] and L. Kurgan[1]

[1] - *Department of Electrical and Computer Engineering, University of Alberta, Canada,* [2] *– Department of Microbiology and Immunology, Weill Cornell Medical College, New York, U.S.A.*
lkurgan@ece.ualberta.ca, mizianty@ualberta.ca

In CASP10 we applied three predictors that were inspired by our MFDp method[1]. These meta-predictors are trained on an updated and larger dataset and include more base disorder predictors. Most importantly, we incorporate two novel ideas, including a new class of input features that are based on alignment to native disorder segments, and a novel set of inputs derived from predicted 3D structure models; the latter are utilized in the biomine_DR_mixed predictor.

Methods

Our predictors work in two steps, where the input chain is first converted into a set of custom-designed numerical features, which are inputted into a logistic regression-based predictor. The design of input features, features selection and training procedure were adopted from the original MFDp method[1]. The source information used to derive the features was extended to include more disorder predictors, including DISOPRED2[2], IUPred[3], MD[4], Norsnet[5], Ucon[6], SPINE-D[7], GlobPlot[8], DisEMBL[9], and PreDisorder[10], and also sequence alignment into a comprehensive dataset of native disordered segments. In the case of the biomine_DR_mixed method we also include structural information derived from 3D models computed with HHPred[11].

Each of the three predictors was trained on a different dataset. The biomine_DR_pdb was trained on the dataset that was used to build ESpritz[12], which annotates NMR-derived disorder; the biomine_DR_pdb_c uses a training dataset with REMARK 465-derived disorder annotations that has low sequence similarity to the CASP9 dataset; the biomine_DR_mixed uses a subset of the latter set.

Unlike a number of other consensus-based disorder predictors, our methods also include other input information sources, such as evolutionary profiles and predicted secondary structure, solvent accessibility, and dihedral angles. The input features include raw values as well as various aggregated values. The biomine_DR_mixed method also uses features generated from the HHpred outputs including evolutionary information from HMM-based substitution matrix, predicted secondary structure, solvent accessibility and predicted B-factor. We empirically evaluated several classifiers, such as logistic regression, linear kernel-based SVM, and RBF network. The best results were obtained using logistic regression, which was adopted in the three predictors.

We also perform post-processing of the predictions from the logistic regression, which is similar as the post-processing that we used during CASP9. Instead of reporting raw predicted probability values for each residue, we aggregate probabilities using the mean value over 5-residues window, and we remove short, up to 2 residues, disordered or/and ordered segments. The optimal size of the removed segments was empirically determined on our training datasets. Finally we optimize thresholds on probabilities to maximize the value of the MCC on the corresponding training datasets. In rare cases where HHpred fails to generate predictions, we replace the corresponding predictions with the predictions from the biomine_DR_pdb_c method.

Results

We evaluated our methods on two datasets, the CASP9 dataset and a PDB-derived dataset that contains all proteins with annotated disorder released after the date when training dataset was obtained; see Table 1. The biomine_DR_mixed method provides the best performing predictions, which suggests that addition of features extracted from the predicted 3D models is helpful. The biomine_DR_pdb method achieves the worst results. This is not surprising as this predictor was trained only on the NMR-derived disorder, which accounts for only about 10% of the CASP9 dataset. Results on the CASP9 dataset are in general lower than for the PDB-derived dataset, and this may be due to lower sequence similarity between our REMARK 465-derived training datasets and the CASP9 set.

Table 1. Results of the evaluation on two test datasets.

Method	CASP9 dataset			PDB-derived dataset		
	MCC	ACC	AUC	MCC	ACC	AUC
biomine_DR_mixed	**0.513**	0.697	**0.879**	**0.651**	0.819	**0.952**
biomine_DR_pdb_c	0.454	0.671	0.850	0.638	0.824	0.938
biomine_DR_pdb	0.346	0.669	0.813	0.449	**0.835**	0.910
prdos2[*]	0.418	**0.754**	0.855	N/A	N/A	N/A
DisoPred3C [**]	0.506	0.670	0.854	N/A	N/A	N/A

[*] – method that achieved the highest ACC and AUC on CASP9

[**] – method that achieved the highest MCC on CASP9

Availability

The MFDp predictor, which is the precursor for these three 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. Ward, J.J., et al. (2004). The DISOPRED server for the prediction of protein disorder. *Bioinformatics* **20**:2138-2139.
3. 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.
4. Schlessinger, A., et al. (2009). Improved disorder prediction by combination of orthogonal approaches. *PloS one* **4**:e4433.
5. Schlessinger, A., et al. (2007). Natively unstructured loops differ from other loops. *PLoS computational biology* **3**:e140.
6. Schlessinger, A., et al. (2007). Natively unstructured regions in proteins identified from contact predictions. *Bioinformatics*, **23**:2376-2384.
7. Zhang, T., et al. (2012). SPINE-D: accurate prediction of short and long disordered regions by a single neural-network based method. *Journal of biomolecular structure & dynamics* **29**:799-813.
8. Linding, R., et al. (2003). GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Research* **31**:3701-3708,
9. Linding, R., et al. (2003). Protein disorder prediction: implications for structural proteomics. Structure **11**:1453-1459.

10. Deng, X., et al. (2009). PreDisorder: ab initio sequence-based prediction of protein disordered regions. *BMC bioinformatics* **10**:436.
11. Söding, J., et al. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research* **33**:W244-248.
12. Walsh, I., et al. (2012). ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics* **28**:503-509.

Model selection using BITS server

A. Roy[1], G. Chopra[1, 2], H. Tiang[1] and R. Samudrala[1]

[1] - *Dept. of Microbiology University of Washington USA,*

[2] - *Dept. of Structural Biology Stanford University USA*

ram@compbio.washington.edu

BITS server is based on the simple premise that the functional sites in protein are more conserved than their global structure, and local structure quality of predicted models can be used as a representative of their structural quality. Models are ranked based on the structural quality of predicted binding site residues. When no binding site is identified, global structural comparisons with a model predicted using multiple templates and scoring with a knowledge-based potential is done to judge the model quality.

Methods

We have a developed a novel computational protocol for model quality assessment using binding site comparisons to judge the quality of the models. The binding site residues in the CASP server models were predicted using the COFACTOR algorithm[1; 2]. COFACTOR scans the query 3D structure against the template library, first based on global structure similarity, followed by a local similarity refinement search on selected templates, with the purpose of filtering out template proteins that do not share binding site similarity with the query protein. During the local structure similarity search, template proteins are scored against the query protein using an innovative structure-sequence similarity measure (BS-score), which is designed to capture both chemical and structural similarity between the query and the template proteins. The template protein, which shared the highest local similarity with the CASP models, was finally used to rank all the models based on their local similarity score (BS-score).

For cases where no confident binding sites was predicted by COFACTOR, we first determined the difficulty of the target ("easy", "medium" and "hard") based on the threading Z-score cutoffs. For "easy" and "medium" targets, global structural similarity[3] to the models predicted by our STRINGS server was used for the ranking. For "hard" cases, template protein which was identified by multiple different threading programs was structurally aligned with the CASP models and global structural similarity score to this template was used for the ranking[4]. When no single template was identified from the consensus, we ranked the models by scoring them with our knowledge-based potential[5,6].

Availability

BITS is available as a web sever at http://cando.compbio.washington.edu/casp/bits/.

1. Roy, A., Yang, J. and Zhang, Y. (2012) COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Res*. **40**, W471-477.
2. Roy, A. and Zhang, Y. (2012) Recognizing protein-ligand binding sites by global structural alignment and local geometry refinement. *Structure*. **20**, 987-997.
3. Zhang, Y. & Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702-10.

4. Zhang, Y. & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* **33**, 2302-9.
5. Chopra, G., Summa, C. M. & Levitt, M. (2008). Solvent dramatically affects protein structure refinement. *Proc Natl Acad Sci USA*. **105**, 20239-20244.
6. Chopra,G., Kalisman,N. & Levitt,M. (2010). Consistent refinement of submitted models at CASP using a knowledge-based potential. *Proteins*. **78**, 2668-2678.

Refinement of server models and refinement category targets using REFINER (extended with mulibody terms) and MODELLER, combined with consensus restraints

M.J. Boniecki[1]

[1] - *International Institute of Molecular and Cell Biology, ul. Trojdena 4, 02-109 Warsaw, Poland*
mboni@genesilico.pl

In CASP10 my participation was mostly focused on refinement. It concerned both refinement of CASP10 server models, sent as regular tertiary structure predictions, and refinement category targets. In case of regular tertiary structure prediction category, the assumption that server models are sufficient approximation of the real target structures was made. Models were evaluated, searched for regions that required refinement and modeled with restraints or, in case of lack of homology, entire structures were subjected for refinement. Refinement was performed using REFINER[1] and MODELLER[2] programs.

Methods

In regular tertiary prediction category CASP10 server models were downloaded and evaluated using MQAPII method[3]. Best scoring models were selected. Manual inspection of templates and alignments that have been used for creation of the models was done. Sets of potential alignments for subsequent targets were taken from Genesilico MetaServer[4]. Further path was similar both for regular TS category and REFINEMENT category. In both cases regions for modeling and restraints were established. It was done by inspecting similarity of the models to their templates, MQAPII's local score and additional organizers' suggestions (if provided). Regions close to the template and well conserved were restrained. In case of non-homology CASP10 server models unrestrained refinement was performed.

Refinement was done using REFINER program, which is an intermediate resolution refinement method based on simplified representation of polypeptide chain, statistical potential and Monte Carlo methods. REFINER represents polypeptide chain as a chain of C-alpha atoms, while the side chains are represented by one or two pseudo-atoms, depending on the size of the side group. REFINER uses quasichemical, orientation dependent contact potential derived from a database of protein crystal structures. Besides regular pair-wise interactions recent version of the program was enriched with terms controlling composition of side group atoms in the vicinity of a given side group atom, which is a step towards multi-body potential terms. Sampling of conformational space was accomplished by usage of asymmetric Metropolis scheme, controlled by Replica Exchange Monte Carlo method. REFINER uses set of moves that allow for conformational change. The recent version was enriched with option of modeling protein oligomers that allows for maintaining symmetry of the system during the entire simulation run. This option was employed for modeling of refinement homo-dimers, where symmetry was derived from the template or provided as organizers' suggestion.

After the simulation run data were subjected for clustering. The lowest energy and most populated conformations were processed further. Full atomic reconstruction was accomplished using REBUILD program from MMTSB package[5]. The last stage of modeling (removal of atom clashes) was done using Modeller. Decoys were clustered and evaluated using MQAPII. Final models were selected based on MQAPII score.

Results

A semi-automated pipeline for protein modeling was created. The method requires a set of structures that approximate the final structure as an input. It also requires alignments and templates that have been used for the generation of input models. Human intervention is mostly needed in the stage of formulating restraints and establishing fragments for refinement.

Availability

REFINER is available publicly (albeit with limited capacity) as a part of the GeneSilico protein modeling toolkit at https://genesilico.pl/toolkit/ .

1. Boniecki,M., Rotkiewicz,P., Skolnick,J., Kolinski,A. (2003). Protein fragment reconstruction using various modeling techniques. *J Comput Aided Mol Des.* **17**, 725-738.
2. Eswar,N., Eramian,D., Webb,B., Shen,M.Y., Sali,A., (2008). Protein structure modeling with MODELLER. *Methods Mol Biol.* **426**, 145-159.
3. Pawlowski,M., Gajda,M.J., Matlak,R., Bujnicki,J.M. (2008). MetaMQAP: a meta-server for the quality assessment of protein models. *BMC Bioinformatics*. **9**, 403.
4. Kurowski,M.A., Bujnicki,J.M. (2003). GeneSilico protein structure prediction meta-server. *Nucleic Acids Res.* **31**, 3305-3307.
5. Feig,M., Karanicolas,J., Brooks,C.L. 3rd (2004). MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology. *J Mol Graph Model.* **22**, 377-395.

RING: a new network method for quality assessment of homology models

Alessandro Masiero[1], Giovanni Minervini[1] and Silvio C.E. Tosatto[1]
[1] - *Dipartimento di Biologia*
Università degli studi di Padova
alessandro.masiero.3@studenti.unipd.it

Protein structures prediction starting from a primary sequence is a very challenging task, as demonstrated by almost two decades of CASP [1] competition. Generating new structures using template based strategies can result in putative structures, especially in side chain placement. The backbone structure prediction is, nowadays, more treated than prediction of the side chains conformations, as tools developed for previous CASP sessions demonstrated. Evaluating the accuracy of the side chains placement can be a crucial point in a homology procedure. In this prediction session, we used the RING[2] (Residue Interaction Network Generator) networks in order to assess the side chain placement and the chemical interactions of our predicted models.
The RING server generates a network for each model, constituted by nodes and edges representing interactions. The represented interactions are Simple Interactions (pairs of C-alpha atoms and close atoms), van der Waals interactions, hydrogen bonds, salt bridges, π-cation and π-π interactions, disulfide bridges and peptide bonds. The output of the RING server is completely compatible with the Cytoscape[3] format. Hence, the generated networks could be visualized and evaluated through this software, and compared with the 3D model structures for a better and comprehensive assessment of the model.

Methods

All models were built using crystal structures available at www.pdb.org database as templates. All the templates were found manually, evaluating each crystal structure by resolution index and visual inspection.

The template search was carried out by means of PDBBlast[4], using the target fasta sequence as input and then choosing the top ranking templates resulting from the resulting alignment. The ranking was based on identity percentage and query coverage percentage. The query sequence was then evaluated by Phyre and Phyre 2 servers[5]. This step was based upon secondary structure evaluation and comparison between the query sequence and the structure database.

The sequence alignment between the target and template fasta files was carried out with ClustalW[6], using blosum matrices and trying to avoid gaps larger than five residues. This was done, where it was possible, in order to avoid poorly guessed coordinates on uncorresponding residues backbone and to maintain the highest rate of secondary structure features of the template.

Once the template was selected and the two sequences were aligned, the models were built through the usage of the HOMER-M[7] server. The side chain placement was performed by SCWRL[8]. The final model was then minimized, in order to remove any van deer Waals and

charge clashes over the generated structure. The energy minimization step was carried out by means of 100ps of molecular dynamics simulation with NAMD[9] as molecular dynamics engine and using the CHARMM 27 force field[10]. The backbone atoms were kept fixed during the minimization step, in order to minimize the loss of secondary structure features. At the end of this step, models were superimposed to templates and were then evaluated by RMSD index to check the secondary structure conservation between template and model.

The final models assessment was achieved by the Ramachandran plot evaluation, verifying that the number of outliers was low. Finally we evaluated the intramolecular interactions with a network representing the model. This was achieved by means of RING, which generated a network for each submitted structure. The network was made by nodes, which represented residues, and edges representing interactions between residues. The RING tool evaluates both the 3D structure and the linear sequence of the target, and provides immediately a wide overview of the residue-residue interactions.

Results

The possibility to evaluate the chemical interactions occurring between the residues of the model in a fast and objective way was short time consuming. It turned out to be a very useful tool, helping us in focusing on key residues and on chemical reliability of the generated models. Visualizing a protein as a network of interactions can highlight structure features that are often not so clear when visualizing just the 3D structure. The residue-residue interactions were both visualized in UCSF Chimera[11] as 3D structure and in Cytoscape as network. Compared to the classical 3D structure evaluation, the simplicity of the RING representation greatly increased our ability to detect unlikely interactions.

Availability

The RING server is available at protein.bio.unipd.it/ring, and the generated output can be opened with Cytoscape. The RINalyzer tool allows to see the pdb 3D structure in UCSF Chimera, simultaneously with the network visualization in Cytoscape.

1. Moult, J., et al. (1995). A large-scale experiment to assess protein structure prediction methods. Proteins 23 (3),2 – 4
2. Alberto J.M. Martin, Michele Vidotto, Filippo Boscariol, Tomás Di Domenico, Ian Walsh and Silvio C.E. Tosatto (2011). RING: Networking interacting residues, evolutionary information and energetics in protein structures. *Bioinformatics*.
3. Michael Smoot, Keiichiro Ono, Johannes Ruscheinski, Peng-Liang Wang, Trey Ideker (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27**(3), 431–432
4. Leszek Rychlewski,1 Lukasz Jaroszewski, Weizhong Li, Adam Godzik (2000). Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Science* **9(2)**, 232-241.
5. Kelley LA and Sternberg MJE (2009). Protein structure prediction on the web: a case study using the Phyre server. *Nature Protocols* **4**, 363 - 371
6. Larkin M., et al. (2007). Clustal W and Clustal X version 2.0, *Bioinformatics* **23(21)**,2947-2948.
7. http://protein.bio.unipd.it/homer/

8. Adrian A. Canutescu, Andrew A. Shelenkov, Roland L. Dunbrack jr. (2003) A graph-theory algorithm for rapid protein side-chain prediction. *Protein Science* **12(9)**,2001-2014
9. James C. Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D. Skeel, Laxmikant Kale, and Klaus Schulten (2005). Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry* **26**, 1781-1802
10. B. R. Brooks, C. L. Brooks III, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus (2009). CHARMM: The Biomolecular simulation Program. *J. Comp. Chem.* **30**, 1545-1615
11. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. (2004). UCSF Chimera -- a visualization system for exploratory research and analysis. *J Comput Chem*. **25(13)**, 1605-12

Modeling protein disorder using CASPITAv2 in CASP9

Ian Walsh, Alberto J. Martin Martin, Tomás Di Domenico and Silvio C. E. Tosatto

Department of Biology, University of Padua, Viale G. Colombo 3, 35131 Padova
ian.walsh@bio.unipd.it

CASPITAv2 is a meta predictor based on (1) CSpritz[1] (see abstract at this CASP), (2) ESpritz[2] (see abstract at this CASP) and (3) PreDisorder (MULTICOM-REFINE in CASP9)[3 4].

Methods
The output probabilities produced by each predictor are simply averaged.

Results
No results were calculated for this method, we are using this round of CASP as a performance evaluator.

Availability
CASPITAv2 is not available.

1. Walsh, I. *et al.* CSpritz: accurate prediction of protein disorder segments with annotation for homology, secondary structure and linear motifs. *Nucleic Acids Res.* **39**, W190–196 (2011).
2. Walsh, I., Martin, A. J. M., Di Domenico, T. & Tosatto, S. C. E. ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics* **28**, 503–509 (2012).
3. Deng, X., Eickholt, J. & Cheng, J. PreDisorder: ab initio sequence-based prediction of protein disordered regions. *BMC Bioinformatics* **10**, 436 (2009).
4. Hecker, J., Yang, J. Y. & Cheng, J. Protein disorder prediction at multiple levels of sensitivity and specificity. *BMC Genomics* **9**, S9 (2008).

Chunk-TASSER server for protein structure prediction in CASP10

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 participated in the CASP structure prediction experiment. For Easy targets (see below), it is a fasrt version of the TASSER-VMT[1] server and for Medium/Hard targets, it is an updated version of the original chunk-TASSER[2]. Chunk-TASSER is much faster than TASSER-VMT, especially for Easy targets.

Method

The Chunk-TASSER server uses an updated version of the SP^3 threading method[3]. 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 targets (SP^3 Z-score >= 6.0), the SP^3 alternative alignment is generated by a parametric alignment method coupled with short TASSER refinement on models selected using knowledge-based scores. The refined top model is structurally aligned to the template to produce the SP^3 alternative alignment. These template models with SP^3 alternative alignments are then grouped into sets containing a variable number of template combinations. For each set, instead of using TASSER as in the TASSER-VMT[1] server, we use MODELLER [4] multiple template modeling to build full-length models. The FTCOM[5] method is used to select five models from the pool of structures generated by MODELLER. For Medium/Hard targets (SP^3 Z-score < 6.0), in addition to threading template models, we also generated full length ab initio models by fragment assembly[6] if the target size is < 200 residues. Threading models and the ab initio models are ranked by FTCOM[5] and the top 20 models are fed into TASSER[7] for refinement. As in our original chunk-TASSER, for Medium/Hard targets, chunk models generated by an 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.

Result

For structures released by Sep 20,2012, chunk-TASSER is only slightly worse than HHpredA(Q) and worse than TASSER-VMT by around 2.5%, results that are consistent with our prior benchmarking. This implies that TASSER refinement performs better than simple multiple template modeling using MODELLER. It is noteworthy that for target T0650, chunk-TASSER is much better than TASSER-VMT (GDT-TS-score: 0.93 vs. 0.73). It might be due to the fact that the majority of FTCOM selected models are worse than the top first model; thus, the TASSER refined model is worse than the top first model.

Availability

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

1. Zhou, H, Skolnick, J. (2012) Template-based protein structure modeling using TASSER^{VMT}. Proteins: Structure, Function, and Bioinformatics. **80**(2):352-361
2. Zhou, H and Skolnick, J. (2007) Ab initio protein structure prediction using chunk-TASSER.. Biophysical Journal. **9**3,1510-8.
3. 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.
4. Sali,A., et.al. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993;234:779--815.
5. Zhou,H and Skolnick, J. (2010) Improving threading algorithms for remote homology modeling by combining fragment and template comparisons. Proteins. **78**, 2041-8.
6. 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.
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.

Prediction of Ligand Binding Sites for Protein Using Isolated FAMSD

Koutarou Yoshiyama[1], Misato Matsuoka[1] ,
Kazuhiko Kanou[2] , Katsuichiro Komatsu[3] , Hideaki Umeyama[1]and Mitsuo Iwadate[1]

[1] – *School of Science and Engineering, Chuo University*
[2] – *National Institute of Infectious Diseases*
[3] – *School of Science, Kitasato University*
a12.px8e@g.chuo-u.ac.jp

Our manual group "chuo-binding-sites" attended ligand bind prediction category (FN) and tertiary structure prediction category (TS) in CASP10. We mainly used the software Isolated FAMSD which automatically executes such some BLAST type homology search programs as PSI-BLAST and the modeling software FAMS-Ligand[1] which makes the protein structure including several ligands indicated without short contacts for distances. In the case of hard targets based upon various sequence alignment programs, we obtained alignments from homology search software SPARKS2, SP3, HHsearch and HMM_BLAST in addition to the alignment programs mentioned above. In order to obtain the answer of ligand binding sites in the ligand sites prediction category (FN), we could make the protein structures including some ligands, and sent the protein coordinates as the results of the tertiary structure prediction category (TS) in CASP10. The protein models made to find the ligand binding sites was distorted according to the binding ligands in family proteins with high sequence homology. Then, we checked the accuracy of the models made from the family proteins including the binding ligands in comparison with the modeling results of our "chuo-fams-server" team in the CASP10 without the consideration of the ligand binding.

Methods

The method used by "chuo-binding-sites" are based on the software Isolated FAMSD which uses sequence alignments obtained from some programs related to the BLAST such as the PSI-BLAST against 95% non-redundant sequences of PDB, ranks alignments by PF_score[2], and builds models so that the atoms of protein and those of ligands don't overlap each other by FAMS-Ligand program. In the Isolated FAMSD system, the source programs are placed in the Linux computer machine which is controlled by the Web Program in Windows, X, Vista and 7 machines. The representative model is selected by four evaluation scores, alignment length, homology, secondary structure information and the CIRCLE score[3], which are calculated for each constructed models. The CIRCLE score is 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.

Next, the representative model selected according to the above algorithm is used as the reference protein superimposed by the protein including some ligand compounds with no short contacts. Thus, the rigidly moved ligands are searched, and the amino acid residues of the protein within 4Åare registered as the ligand binding sites. The representative protein model changed in the protein-ligand interactions by the FAMS-Ligand program is determined as the 3-dimensional

protein structure proposed in the CASP10.

Moreover, in the case of the difficult target in the sequence alignment, some alignment programs such as SPARKS2, SP3, HHsearch and HMM_BLAST are used. The determination of the ligand binding sites and the 3-dimensional structures is similarly performed with the method mentioned above.

Results

The number of the PDB coordinates corresponding to the answers for the target query sequences obtained from CASP10 organizers by September 6, 2012 are 34. The GDT_TS values for the models proposed by us were calculated. We compared the GDT_TS values of the "chuo-binding-sites" team with those of our "chuo-fams-server" team. The accuracy of the models made from the family proteins including the binding ligands is comparable for the modeling results of our "chuo-fams-server" team in the CASP10 without the consideration of the ligand binding. Accordingly, the results in which the binding sites are searched may be reliable.

Next, there were 27 answer PDB files including ligands out of 34 answer PDB files. Here, if the atoms of ligands exist within 4.0Å from the amino acid residues of the target protein, the amino acid residues were defined to be the ligand binding site. Thus, the value of 4.0Å was also used to determine the binding site for the CASP10 target models. Figure 1 and 2 show the number of applicable protein targets against the accuracy rates of the binding sites composed of amino acid residues for all the models of 27, and the models over the GDT_TS value 30, respectively.

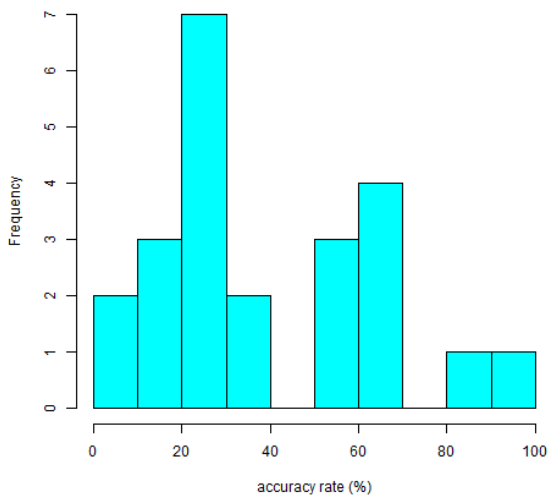


Fig.1 All the models of 27
(The number of the accuracy rate 0% is 4.)

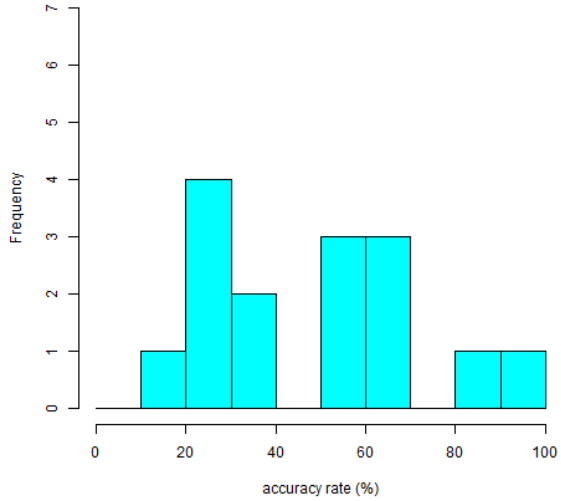


Fig.2 The models over the GDT_TS value 30
(The number of the accuracy rate 0% is 3.)

1. Umeyama H, Iwadate M., FAMS and FAMSBASE for Protein Structure. In Current Protocols in Bioinformatics. John Wiley & Sons, Inc. (2002)
2. Iwadate M, Kanou K, Terashi G, Umeyama H, Takeda-Shitaka M., Chem. Pharm. Bull., 58, 1-10 (2010)
3. Terashi G, Takeda-Shitaka M, Kanou K, Iwadate M, Takaya D, Hosoi A, Ohta K, Umeyama H., *Proteins*, 69, Sp.8, 98-107 (2007)

Tertiary structure prediction of chuo-fams team by the consensus method composed of GDT_TS and secondary structure arrangement under the check of the packing free energy using STAGE2 server models

Koutarou Yoshiyama, Wataru Sato , Misato Matsuoka ,
Takahiro Miyashita , Hideaki Umeyama , Mitsuo Iwadate

School of Science and Engineering, Chuo University
a12.px8e@g.chuo-u.ac.jp

Our manual group "chuo-fams" attended tertiary structure prediction category (TS) in CASP10. Tertiary structure prediction of "chuo-fams" team by the consensus method composed of GDT_TS and the sequential arrangement of the secondary structure under the check of the packing free energy using STAGE2 server models was performed.

We notice that the results of our "chuo-fams" team are compared with those of the "Zhang-Server" (http://zhanglab.ccmb.med.umich.edu/casp10/11.html), because, as shown in the URL, the "Zhang-Server" ranks first in the Cumulative Score of 51 targets of CASP10 until September 22, 2012. The URL table is ranked by TM-score of the first model. We are very interesting to compare the "Zhang-Server", because our "chuo-fams" wants to get more superior results than any server teams.

Methods

For the tertiary structure (TS) category, we performed our manual prediction using the following four scores for the selection of first ranking model in the STAGE2 files of the CASP10 site, which had about 150 server models.

1. Cα atom consensus
   The Cα coordinates are used in the calculations of the GDT_TS value with each other team. The GDT_TS values are used in the consensus method for each server model.
2. Secondary structure consensus
   The sequential arrangement of the secondary structure is compared, and it is used in the consensus method for each server model. The secondary structures are obtained from the DSSP[1] calculations.
3. CIRCLE[2]
   The score in the packing of amino acid residues based on physiochemical free energy is calculated from the CIRCLE program.
4. ss_score[3]
   The score representing the rate of secondary structure identity between PSIPRED[4] prediction of the target protein sequence and STRIDE[5] judgment of the model protein is used.

The weights of these scores depend on the difficulty of each target. For easy targets, we attached importance on the consensus scores, but on the other hand, we attached importance on the absolute scores such as CIRCLE and ss_score for hard targets.

Results

The number of the PDB codes corresponding to the answers for the target query sequences obtained from CASP10 organizers by September 6, 2012 are 34. The GDT_TS values for the models proposed by us were calculated. We compared the GDT_TS values of the "chuo-fams" team with those of the "Zhang-Server" team. Figure 1 shows the plot of the "Zhang-Server" team against the "chuo-fams" team. This figure indicates that the consensus method of the "chuo-fams" team is useful. Accordingly, the results in which our consensus method in the "chuo-fams" may be reliable are obtained.

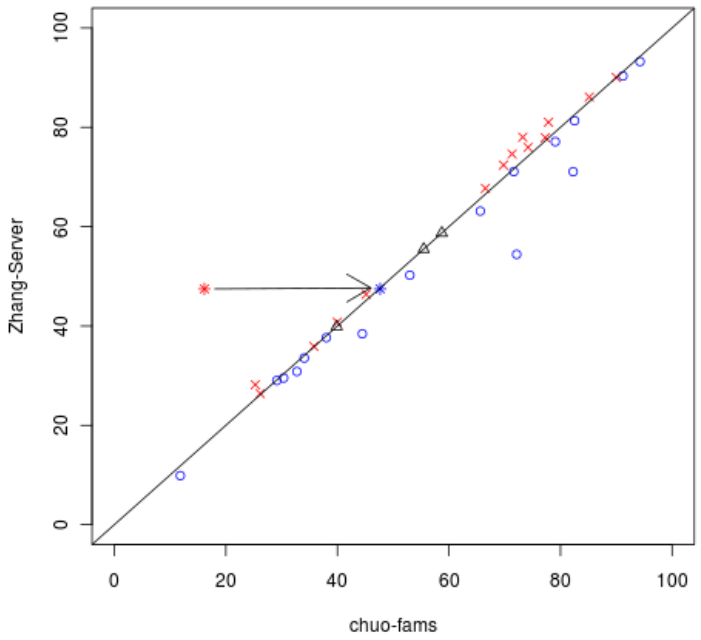


Fig.1 The GDT_TS plot of the "Zhang-Server" against the "chuo-fams".

Here, the superior result of the "chuo-fams" is described by the circle, and that of the "Zhang-Server" is described by the cross. The triangle means the same GDT_TS values. There was one careless miss in the "chuo-fams" team, though we didn't have such a careless miss for any other targets. If we have selected according to the model written in the note, the result would change to the place where the arrow indicates in this figure.

1. Cabsh W, Sander C, *Biopolymers*, 22, 2577-2637 (1983).
2. Terashi G, Takeda-Shitaka M, Kanou K, Iwadate M, Takaya D, Hosoi A, Ohta K, Umeyama H. *Proteins*, 69, Sp.8, 98-107 (2007)
3. Arai M, *Construction of the Function for Protein Structure Prediction and the Homology Modeling*, Chuo University (2011)
4. Jones, D.T., *J. Mol. Biol*. 292:195-202 (1999)
5. Heinig, M., Frishman D, *Nucl. Acids Res.*, 32, W500-2 (2004)

3-Dimensional models created by the chuo-fams-consensus team using three consensus methods and the FAMS protein modeling program

Wataru Sato, Hideaki Umeyama, Mitsuo Iwadate

School of Science and Engineering, Chuo University
a12.4eh8@g.chuo-u.ac.jp

Our human team "chuo-fams-consensus" attended TS(3D atomic coordinates prediction) category in CASP10. For attending the CASP10, we use three kinds of consensus methods based on the experiences of past CASPs.

We use three information of the sequential arrangement of secondary structure, the 3-dimensional arrangement for the conserved regions of the secondary structure, and the conservation of 3-dimensional structure based upon the positions of Cα in the models.

Methods

First consensus method is performed among many server models in the STAGE2 using the score in the sequential arrangement of the secondary structure for other server models. We used the DSSP program to determine the secondary structure of each amino acid residue for all the server models in the STAGE2.

We count the number of the amino acid residue in relation to the agreement of the secondary structure in the subject alignment, and calculate the protein score based upon the number of the agreement of the secondary structure. The ratio of the secondary structure agreement within one region of the secondary structure is defined to be over 60%.

Secondly, we determine the 3-dimensional arrangement of the secondary structure in a server model, and the consensus method is applied in the calculation of the difference of the 3-dimensional arrangement between two server models. The secondary structure consists of amino acid residues over 4 for the same secondary structure. The ratio of the secondary structure agreement of two corresponding amino acid residues should be over 60% in the sequence alignment between the subject two protein models.

The equation to calculate the similarity for the 3-dimensional arrangement of the secondary structures is as follows,

$$\text{Score} = 10 \hat{} (-\text{RMSD}) \times 1000$$

Thirdly, the conservation of 3-dimensional structure based upon the positions of Cα in the two server models is noticed. The GDT_TS value is used as the indication of the conservation of 3-dimensional structure based upon the positions of Cα. Although we use the regions of only alpha-helix and beta-sheet structure in the first and second consensus process, other regions such as loops against the above two secondary structures are included in the third consensus process.

We used the score which we linearly added the consensus scores of above three methods as the final consensus score of the model.

Results

We notice that the results of our “chuo-fams-consensus” team are compared with those of the Zhang-Server (http://zhanglab.ccmb.med.umich.edu/casp10/11.html), because, as shown in the URL(Uniform Resource Locator), the Zhang-Server ranks first in the Cumulative Score of 51

targets of CASP10 until September 22, 2012, though our chuo-fams-server ranks 24. The URL table is ranked by TM-score of the first model. We are very interesting to compare the Zhang-Server, because our “chuo-fams-consensus” team want to get more superior results than any server teams.

The number of the PDB coordinates corresponding to the answers for the target query sequences sent from CASP10 organizers are 34 until September 6, 2012. The GDT_TS values for the models sent to the organizer by us were calculated. We compared the GDT_TS values of the “chuo-fams-consensus” team with those of the “Zhang server” team. The plot of the “Zhang server” team against the “chuo-fams-consensus” team was described. A little bit superior results of the “chuo-fams-consensus” are described in the above plot. Thus, it is indicated that the consensus method of the “chuo-fams-consensus” team is useful.

Accordingly, the results in which our consensus method in the “chuo-fams-consensus” team may be reliable may be obtained.

Availability

1. DSSP：Kabsch W, Sander C., Biopolymers. Dec;22(12):2577-637. (1983)
2. PSIPRED：Jones, D.T., *J. Mol. Biol. 292*:195-202 (1999)
3. FAMS : Ogata K, and Umeyama H., J Mol Graph Model. Jun;18(3):258-72,305-6.(2000)

3-Dimensional models created by Chuo-fams-server team using the FAMS program and some consensus methods

Wataru Sato, Hideaki Umeyama, Mitsuo Iwadate

Graduate School of Science and Engineering, Chuo University
a12.4eh8@g.chuo-u.ac.jp

Our server team, "chuo-fams-server", attended TS(3D atomic coordinates prediction) category in CASP10. We constructed consensus methods and scoreA(score of model accuracy) which consists of the sequence alignment length, the genetic similarity score among 20 amino acid residues obtained from the table of pair points in the BLOSUM62 matrix, the ratio of the agreement of the secondary structure between the predicted secondary structure of the target protein and the secondary structure of the referred PDB protein in the sequence alignment . We used the FAMS as the homology modeling program and PSI-BLAST, HHsearch, SPARKS2, SP3, HMMER and HHM_BLAST as the sequence alignment programs.

In order to increase the accuracy, the GDT_TS value, of the model sent to the CASP10 organizer as the predicted model, we devised the scoring function for the models made in our laboratory. We used also the equation made from the scoreA developed in the CASP9, and some consensus methods mentioned in the Method are applied.

Methods

From the secondary structure point of view, first, the consensus method is applied to the server models in the STAGE2 step. The conservation of the secondary structure among server models from different teams is noticed, and the ratio of the conservation is calculated for pair protein models in the STAGE2 to calculate the consensus score. We used the DSSP program to determine the secondary structure.

From the 3-dimensional structure point of view, secondly, the consensus method is applied to the server models in the STAGE2 step. The 3-dimensional agreement of C alpha atoms among the STAGE2 server models is noticed, and the GDT_TS value indicating the agreement of C alpha atoms is used in the consensus method. Thus, the consensus score of the whole shape of the modeled protein is obtained.

From the sequence alignment point of view, thirdly, we made the score of the sequence similarity which is called "scoreA". The scoreA for one server model consists of the length of sequence alignment using alignment programs, the sum of pair points from BLOSUM62 for the corresponding amino acid residues in the sequence alignment, and the sequential sum of the score for the agreement of the secondary structures between two server models.

Thus, we explain the homology modeling system, in which we incorporated some methods mentioned above. As the first step, we get credible sequence alignments between the target protein and the PDB protein using various alignment programs.

As the second step, we create the 3-dimensional structure model for each alignment using the FAMS program. As the third step, we screen the candidate models using the CIRCLE program calculating the packing free energy in the solution and the length of the sequence alignment. As the 4-th step, we calculate some consensus scores and the score for the scoreA. As the 5-th step, in order to get the final conclusion for the model sent to the organizer, the final

score is calculated using the equation obtained from the training for the CASP8 targets using the maximization of the sum of the Z-score.

Results

We notice that the results of our “chuo-fams-server” team are compared with those of the Zhang-Server(http://zhanglab.ccmb.med.umich.edu/casp10/11.html), because, as shown in the URL(Uniform Resource Locator), the Zhang-Server ranks first in the Cumulative Score of 51 targets of CASP10 until September 22, 2012, though our "chuo-fams-server" ranks 24. The URL table is ranked by TM-score of the first model.

We are very interesting to compare the Zhang-Server, because our “chuo-fams-server” team want to get corresponding results to the "Zhang-Server" team. The number of the PDB coordinates corresponding to the answers for the target query sequences sent from CASP10 organizers is 34 until September 6, 2012. The GDT_TS values for the models proposed by us were calculated. We compared the GDT_TS values of the “chuo-fams-server” team with those of the “Zhang server” team. The plot of the “Zhang server” team against the “chuo-fams-server” team indicates that the results of the “Zhang server” are superior to our ones. However, the plot indicates that the “chuo-fams-server” team is useful and reliable to create the protein model.

1. Ogata K, and Umeyama H. FAMS . J Mol Graph Model. Jun;18(3):258-72,305-6.(2000)
2. Arai M. Construction of the Function for Protein Structure Prediction and the Homology Modeling, Chuo University (2011)

Study on 3-Dimensiional repacking between such secondary structures as alpha helix and beta sheet in the homology modeling process

Takahiro Miyashita, Wataru Sato, Koutarou Yoshiyama, Hideaki Umeyama, and Mitsuo Iwadate

School of Science and Engineering, Chuo University
a12.hbag@g.chuo-u.ac.jp

In a little bit hard homology modeling, generally, the ratio of sequence similarity is under 20% between the modeling target protein and the 3-dimensionally analyzed reference protein. Even if the whole protein structure seems to be similar, the 3-dimensional arrangement among neighboring secondary structures will be positionally misaligned between the target protein and the reference protein because of the low sequence similarity which brings the change of the shape of the secondary structure. In the FAMS program, which we are using as the homology modeling soft, the positions of the secondary structures do not move or misalign largely due to the constraint energy not to move from the referred main chain positions according to the training research among the family proteins over the sequence similarity of about 35%. We are thinking that the introduction of the misalignment of the secondary structure is very important to create a homology model of the low sequence similarity. In the CASP10 contest, thus, we attended the competition as the teams of “chuo-repack-server” and “chuo-repack” in order to resolve the misaligned problem of the secondary structure in the hard homology modeling. In order to select the subject model proteins, we used the server models and the manual ones. As the former models, we used some models from "chuo-fams-server" and "chuo-binding-sites" teams. And, as the latter models, we used some models from "chuo-fams" and "chuo-fams-consensus" teams.

Methods

Let's explain how to select the repack protein. First, we identified the regions of the secondary structures such as alpha helix and beta sheet using the program called "ksdssp". Since we need to misalign only the secondary regions defined by us, we deleted the regions including the loops in the model. . From the visual point of view, we selected the pair secondary structures, which seem to be weak in the interaction with a slightly large distance. And we moved the relative positions of the secondary structures in the calculations of the docking score. In order to reattach the loop regions excluded in the process of the 3-dimensional misalignment, we executed the FAMS program, which reforms the whole protein structure using the repacked secondary structure regions and the loop regions corresponding to the reference protein in the homology modeling.

Explanation of docking program between two secondary structures. We define each of all the secondary structures about whether the secondary structure should be subject or not. The rigidly transferred and rotated secondary structure is named to be "str1" and, the fixed secondary structure is named to be "str2". For the atoms composing the "str2", each distance between two grid places around the "str2"are determined to be about 1 Å, and the zero value is given to those grid places. And the grid places within 3Å from the atoms of the "str2" obtain minus fifteen points. After that, for the grid places having the zero value in the edge of zero regions between the zero regions and the minus fifteen region of grid places, the points are changed and reset to be one point from the zero point. Next, the gravity center of the "str1" is fitted to a grid place belonging to the "str2", and the points corresponding to the grid places contained within 3Å from

the atoms in the "str1" are added. And this total score is defined as an index for the surface area of the contact region between the "str1" and the "str2". Then, we change the corresponding grid place. The gravity center of the "str1" is fitted to other grid place belonging to the "str2", and the points corresponding to the grid places contained within 3Å from the atoms in the "str1" are added again. On the other hand, the Euler's three angles around the gravity center of the "str1" are changed in turn less than 15 degrees. Again, similarly the maximum score is calculated for many small rotations of the "str1". Thus, the top score in the two processes of the gravity center of the "str1" fitting to the "str2" grid place and the rotation of the Euler's three angles around the gravity center of the "str1" is calculated to determine the interacting structure between two secondary structures.

Results

In the CASP10 contest, the repacking calculations were performed for the 25 target in the "chuo-repack-server" and the 24 targets in the "chuo-repack". Using the PDB coordinates published until September 12, 2012, the GDT_TS value were calculated to compare with the GDT_TS values of the models used as the starting structure. The results are shown in Figures 1 and 2. Two figures are based upon the "chuo-repack-server" and "chuo-repack" teams, respectively. For almost models, the repacking process does not work to improve the accuracy of the starting models. Conclusively, we should revise the docking program between two secondary structures to become more accurate.

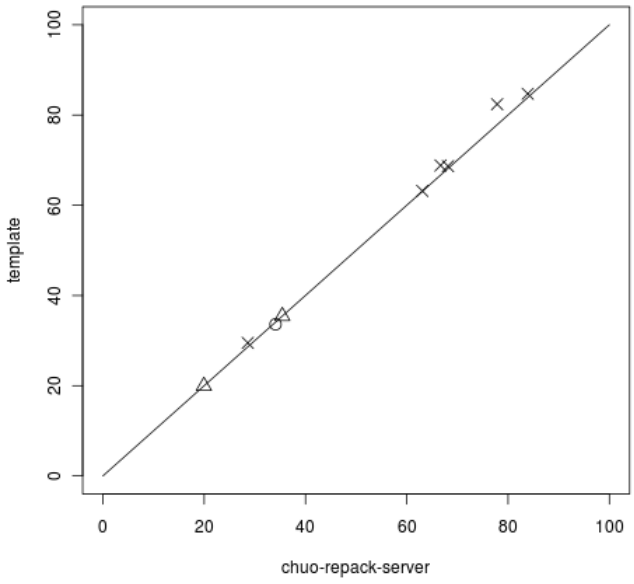


Figure 1. Show the GDT_TS plot of the "template" against the "chuo-repack-server". The superior results of the "chuo-repack-server" are described by the circle, and that of the "template" is described by the cross. The triangle means the same GDT_TS values.

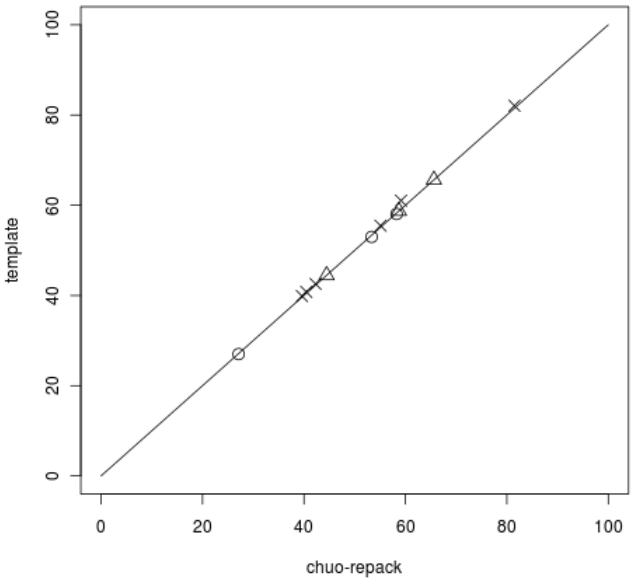


Figure 2. Show the GDT_TS plot of the "template" against the "chuo-repack". The superior results of the "chuo-repack" are described by the circle, and that of the "template" is described by the cross. The triangle means the same GDT_TS values.

1. Ogata K, and Umeyama H., J Mol Graph Model. Jun;18(3):258-72, 305-6. (2000).
2. Arai M, *Construction of the Prediction Function of Protein Structure and the Homology Modeling System*, Chuo University, (2011).

Models from contacts

J. Rodriguez[1], P. Maietta[1], D. Juan[1], I. Ezkurdia[1], F. Abascal[1], A. Valencia[1], M.N. Wass[1] and M.L. Tress[1]

[1] - CNIO *(Spanish National Cancer Research Centre), Madrid, Spain*
mtress@cnio.es

Here we describe the protocols used by the CNIO group for the 3D structure prediction, contact prediction and function prediction experiments during the 10th edition of CASP. The 3D structure predictions were based on our contact predictions, while the human function prediction method combined the predictions of two servers, *firestar*[1] and 3DLigandSite[2].

Methods

Structure and contact prediction

We have shown that the information contained in predicted residue–residue contacts can aid 3D model prediction[3]. The power of predicted contacts to help in 3D structure prediction was surprising because the residue–residue contact prediction methods used in the study had very poor accuracy.

Recent work[4] has suggested that contact predictions can be drastically improved given the right conditions. Unfortunately these conditions are rarely met, in particular for CASP targets. However, we have recently developed a similar contact prediction method that requires alignments with fewer sequences to make predictions.

Our methodology follows the rational of ContextMirror, an approach intended to detect pairs and groups of co-evolving proteins in which covariation remains significant after removing the contribution of others proteins in the alignment[5]. This idea has been adapted to contact prediction starting from the correlated mutations between every pair of residues as calculated by Göbel et al[6]. We use the correlation matrix to define the co-evolutionary profile of each position as a vector containing the correlation values of its position with every other position of the alignment. The similarities between every pair of co-evolutionary profiles are quantified as their Pearson's correlation coefficient. In order to evaluate inter-position specific co-evolution, partial correlation coefficients are calculated for pairs of positions and every possible third position.

Contact pairs were generated by ContextMirror, from the server models and from the binding sites predicted by *firestar* and were used to discriminate between the decoys generated by server groups and generated in house. In CASP10 we predicted all human targets.

Function prediction

For the function prediction experiment we combined the predictions of 3DLigandSite and *firestar*. Although 3DLigandSite had made predictions for all targets (as a structure-based function prediction method), the automatic *firestar* server was only able to detect the required homology in 44 targets. However, evidence from the extended *firestar* alignments and confirmatory evidence from 3DLigandSite allowed us to add a further 16 confident predictions. In almost all cases these were substrate binding sites.

1. Lopez, G, Maietta, P, Rodriguez, JM, Valencia, A and Tress, ML (2011). *firestar* –

advances in the prediction of functionally important residues. *Nucleic Acids Res.*, **39** W235-241.
2. Wass, MN, Kelley, LA and Sternberg MJ (2010). 3DLigandSite: predicting ligand-binding sites using similar structures. *Nucleic Acids Res.*, **38** W469-473.
3. Tress, ML and Valencia, A (2010). Predicted residue-residue contacts can help the scoring of 3D models. *Proteins*, **78**, 1980-1991
4. Morcos, F, Pagnani, A, Lunt, B, Bertolino, A, Marks, DS, Sander, C, Zecchina, R, Onuchic, JN, Hwa, T and Weigt M (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci U S A*, **108** E1293-E1301.
5. Juan, D, Pazos, F and Valencia, A (2008). High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proc Natl Acad Sci U S A*, **105**, 934–939.
6. Göbel, U, Sander, C, Schneider, R and Valencia, A (1994). Correlated mutations and residue contacts in proteins. *Proteins*, **18**, 309–317.

Protein-ligand binding sites prediction using COFACTOR

Ambrish Roy

Department of Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw Ave, Ann Arbor, MI 48109, USA
ambyroy@uw.edu

During CASP10, both our automated server (as "COFACTOR") and human group (as "COFACTOR_HUMAN") used the COFACTOR[1] algorithm and BioLiP database[2], for ligand binding site residue predictions. Both the groups used the I-TASSER 3D models which were built using multiple templates as well as single template models. In addition to the 3D structure predictions, COFACTOR algorithm consists of two key steps for binding site identifications: (a) template identification; (b) local similarity refinement search on selected templates.

Template identification

Template proteins were identified by matching the 3D models with a template proteins in the BioLiP database using the structural alignment program TM-align[3]. All the template proteins having non-random structural similarity, i.e. TM-score >0.3[4], to query structure were used as an input for the next step of local similarity refinement search for binding site residue identification.

Local structural refinement

A binding site refinement search is performed on the selected templates, with the purpose of filtering out templates that even though are globally similar, they do not share binding site similarity with the query protein. This local similarity refinement involves the following steps:

a. *Generation of candidate binding site motifs in query*: Conserved residues in query protein are identified based on Z-score of Jensen–Shannon divergence score[5] and residues with Z-score > -0.2 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 the 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 for increasing 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[6] similar to that used in TM-align[3], where the score for aligning *i*th residue in query and *j*th residue in template is given by the sum of BLOSUM-62 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, and finally the region which gives the best binding site score (BS-score) is selected as the identified binding site in the query and the residues aligned with known binding site residues in the template as binding site residues.
e. *Clustering of binding sites*: Predicted binding sites are clustered based on their spatial distance and chemical similarity of template ligand to generate the final predictions.

Availability

The algorithm is implemented on both I-TASSER (http://zhanglab.ccmb.med.umich.edu/I-TASSER) and COFACTOR (http://zhanglab.ccmb.med.umich.edu/COFACTOR) servers, where the I-TASSER server starts from a target sequence and the COFACTOR server from a 3D structure model of the target.

1. Roy, A. & Zhang, Y. (2012). Recognizing protein-ligand binding sites by global structural alignment and local geometry refinement. *Structure* **20**, 987-97.
2. Yang, J., Roy, A. & Zhang, Y. (2012). BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res*, (submitted).
3. Zhang, Y. & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* **33**, 2302-9.
4. Xu, J. & Zhang, Y. (2010). How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **26**, 889-95.
5. Capra, J. A. & Singh, M. (2007). Predicting functionally important residues from sequence conservation. *Bioinformatics* **23**, 1875-82.
6. 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.

Fully Automated Protein-Ligand Contact Prediction

D.W.A. Buchan[1] D. Cozzetto[1] and D.T. Jones[1]

1 – Bioinformatics Group, Department of Computer Science, University College London, Gower St., London, WC1E 6BT, United Kingdom
d.buchan@cs.ucl.ac.uk
URL: http://bioinf.cs.ucl.ac.uk/psipred/

This method tests a novel, fully automated predictor of protein-ligand contacts based on fold recognition recently developed by the Bioinformatics Group at UCL. This was recently made available as part of the PSIPRED webserver.

Methods
The method is an 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 [1]. We then identify each template structure's ligand-interacting residues using the annotations from PDBSum [2] and the Catalytic Site Atlas [3]. Using the annotations in SwissProt/Uniprot [4], in the Binding MOAD database [5] we identified the set of biologically valid ligands bound to each template and we discard any "non-valid" ligands from each template structure.

With the valid ligand contacts and the initial pair-wise alignments in hand, interacting residue coordinates were mapped from each template onto the target sequence to build a pseudo-alignment of contacts .Contact positions are predicted in the target sequence using a Support Vector Machine approach. Each residue is analysed in turn by taking in to account the number of consensus contacts for the residues (contact propensity) and some additionaly contact metrics within a window of plus and minus 5 residues (these including average number of contacts per position in the window and maximum number of contacts in that window). The SVM was previously trained and benchmarked using a large set of known PDB enzymes and shown to have good ability to discriminate ligand and non-ligand binding residues and modest performance correctly assigning contacts. The CASP10 entry would constitute a "real-world" test of sequences without prior knowledge of the target's status as enzymes.

Results
We submitted predictions for all targets that could find a sufficient number of homologues by fold recognition, without attempts to recognise if the target sequence was an enzyme.

Availability
This predictor can be accessed from the following URL as part of the pGenTHREADER analysis: http://bioinf.cs.ucl.ac.uk/psipred

1. 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.

2. Laskowski, R.A. (2009) PDBsum new things, Nucleic Acids Res, 37, D355-359.
3. 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.
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.

Multiplayer online game-based homology and ab-initio modeling

O. Colluphid and other Foldit players[1]

[1] - *Worldwide*
cfc@folditcontenders.com

Models were constructed and refined using Foldit, the online multiplayer game available at http://fold.it. CASP10 targets shorter than 170 residues were given to the playing community at Foldit as puzzles to solve. The Contenders are a human based subgroup within that community. They also drink a lot of tea.

Methods

Foldit uses the Rosetta protein modeling software package[1] and allows players to modify and visualise protein structures in real time[2]. Foldit players are provided with tools that allow them to 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 minimisation, fragment insertion) are Rosetta optimisations modified to suit direct protein interaction and simplified to run at interactive speeds. Each CASP10 puzzle was typically accessible to Foldit players for 8-9 days.

For CASP10 targets shorter than 170 residues in the "All Groups" category, two different Foldit puzzles were given to the players. One puzzle started from an extended chain, with alignments to known templates taken from the RAPTOR[3], SPARKS[4], and HHsearch[5] servers provided. 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 in-game tools listed above. For the second puzzle, models were constructed using the QUARK[8] and Zhang[7]- Server predictions. These server models were initially minimised using Rosetta and then given as starting points for the Foldit players to refine. This same protocol was used for CASP10 targets in the "Refinement" category, where server models were first minimised with Rosetta before being given to the Foldit playing community. Foldit players were provided with secondary structure predictions, generated by the SAM-T08 server[6], in the form of a sequence logo for all CASP10 puzzles.

Quality and ranking of individual models was determined entirely by the Rosetta full-atom energy ranking. Contenders submissions were then further judged and selected on merits of diversity, compactness and optimal side-chain interaction.

Availability

Foldit is available via the Rosetta Commons at http://tinyurl.com/academic-foldit

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. *Methods in enzymology* 487, 545-74.
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. Peng,J. & Xu,J. (2009) Boosting Protein Threading Accuracy. *Research in Computational Molecular Biology (RECOMB)*, **5541**, 31-45.
4. Yang,Y., Faraggi,E., Zhao, H. & Zhou,Y. (2011) Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of the query and corresponding native properties of templates. *Bioinformatics* **27**, 2076-2082.
5. Söding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics.* **21**(7):951-60.
6. Karplus,K. (2009) SAM-T08: HMM-based Protein Structure Prediction. Nucleic Acids Research. **37**(2): W492-7.
7. Zhang,Y. (2008) I-TASSER server for protein 3D structure prediction. BMC Bioinformatics, vol **9**, 40.
8. Xu,D. & Zhang,Y. (2012) Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins **80**, 1715-35.

Physics-based protein-structure prediction with the coarse-grained UNRES force field

Yi He[1], Adam K. Sieradzan[2], Paweł Krupa[1,2], Magdalena Mozolewska[1,2], Tomasz Wirecki[2], Shalom Rackovsky[1,3], Adam Liwo[2], Stanisław Ołdziej[1,4], 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]-Department of Biomathematical Sciences, Mount Sinai School of Medicine, One Gustave L. Levy Place, New York, N.Y. 10029,[4] – 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 the following four steps. First, a UNited-RESidue force field (UNRES) was employed to carry out Multiplexed Replica Exchange Molecular Dynamics (MREMD)[1] for target proteins. Second, based on the MREMD simulation results, the Weighted-Histogram Analysis Method (WHAM) analysis was used to calculate relative free energy of each structure of the last slice of the MREMD simulation; the respective procedure is described in ref. 2. Third, cluster analysis was employed to cluster the structures from the MREMD simulation. Five clusters with lowest free energies were chosen as prediction candidates. The conformations closest to the respective average structures corresponding to the resulting clusters were converted to all-atom structures[3,4] in the last stage of the approach to produce the models which were subsequently submitted.

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 cluster-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. In this CASP exercise we also introduced correlation terms that couple the backbone and side chain local-interaction energies. The effective energy function depends on temperature and has been parameterized to reproduce structure and thermodynamics of selected training proteins. [2,5]

To obtain prediction, MREMD simulations were run with the parallelized UNRES code available at www.unres.pl. 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.

Availability
The UNRES package to perform coarse-grained simulations is available at http://www.unres.pl

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 α and α+β proteins. *Journal of Chemical Theory and Computation.* **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. *The Journal of Physical Chemistry B* **111**, 260-285.

3. Kazmierkiewicz, R., Liwo, A. & Scheraga, H. A., (2002). Energy-based reconstruction of a protein backbone from its α-carbon trace by a Monte Carlo method. *Journal of Computational Chemistry* **23**, 715-723
4. Kazmierkiewicz, R., Liwo, A. & Scheraga, H. A. (2003). Addition of side chains to a known backbone with defined side-chain centroids. *Biophysical Chemistry* **100**, 261-280.
5. 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. *Journal of Computational Chemistry* **30**, 2127-2135
6. McGuffin, L.J., Bryson, K. & Jones D.T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics* **16**, 404-405.

Modeling protein disorder using CSpritz in CASP9

Ian Walsh, Alberto J. Martin Martin, Tomás Di Domenico and Silvio C. E. Tosatto.

Department of Biology, University of Padua, Viale G. Colombo 3, 35131 Padova
ian.walsh@bio.unipd.it

We retrained our server CSpritz[1] on recent PDB structures. The server is a substantial redesign of our previous method Spritz[2]. The 'C' stands for Combination, meaning we utilize a combination of machine learning and modeling techniques. Structural templates are incorporated in two modules. While the other module is strictly pattern detection (i.e. ab initio). Benchmarking performed on the now two year old CASP9 data indicated CSpritz would have ranked consistently well with other methods.

Methods

CSpritz constitutes three separate modules. Previously, we proved that the modules find different disorder patterns (maximum Pearson correlation coefficient 0.59). This is an important characteristic when doing a combination based approach[3].

Module one: Support Vector Machines (SVMs)[4] were used to find disordered patterns given a local sequence window, predicted secondary structure and PSI-BLAST[5] multiple sequence alignments. In addition, if PSI-BLAST returned PDB entries which contain structure ("non-disorder") the resulting output probability was scaled accordingly. The disorder dataset was not updated for this round of CASP and is identical to the original disorder definition[2]. However, the structural templates used for post-filtering were found on a more recent PDB.

Module two: The same SVM was constructed with the addition of solvent accessibility and structural templates as input. The definition of disorder is also slightly different relative to the other modules. The sequence of each chain in the PDB SEQRES records is aligned with DSSP sequence in the PDBFinderII[6] database. Missing atoms retrieved from this alignment are considered disordered. Only high quality structures were considered (<2.0 Å resolution). Moreover, we used a publicly available dataset[7]. Structural templates formed an input feature whereas module one had it as an output scaling factor.

Module three: see ESpritz group at this CASP. It is important to mention, for this module, NMR based predictions were executed when a target was determined as such.

After careful experimentation a simple average of the probabilities was observed to be the best mode of combination. All predictions were solely computational. Benchmarking was done on the previous CASP round where all data was independent from training.

Results

On CASP 9 targets, our calculations show CSpritz achieved (sensitivity+specificity)/2=76.8 and an Area Under the receiver operator Curve (AUC) 84.3. This is comparable to the best methods. Although it is a meta/combination based approach, all modules were carefully designed, in-house, to be different and to incorporate various sources of information. We anticipate changes in the performance depending on the structural templates found and the amount of NMR structures at this CASP round.

Availability

The server, together with help and methods pages including examples, are freely available at URL: http://protein.bio.unipd.it/cspritz/. An important addition to the server includes the detection of disordered linear motifs from ELM[8], secondary structure and homology information. The server also incorporates Disprot based predictions which are generally longer disorder patterns.

1. Walsh, I. *et al.* CSpritz: accurate prediction of protein disorder segments with annotation for homology, secondary structure and linear motifs. *Nucleic Acids Res.* **39**, W190–196 (2011).
2. Vullo, A., Bortolami, O., Pollastri, G. & Tosatto, S. C. E. Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines. *Nucleic Acids Res.* **34**, W164–168 (2006).
3. Sollich, P. & Krogh, A. *Learning with ensembles: How over-fitting can be useful*. (1996).
4. Cortes, C. & Vapnik, V. Support-Vector Networks. *Mach. Learn.* **20**, 273–297 (1995).
5. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
6. Hooft, R. W. W., Sander, C., Scharf, M. & Vriend, G. The PDBFINDER database: a summary of PDB, DSSP and HSSP information with added value. *Comput Appl Biosci* **12**, 525–529 (1996).
7. Cheng, J., Sweredoski, M. J. & Baldi, P. Accurate Prediction of Protein Disordered Regions by Mining Protein Structure Data. *Data Min. Knowl. Discov.* **11**, 213–222 (2005).
8. Dinkel, H. *et al.* ELM--the database of eukaryotic linear motifs. *Nucleic Acids Research* **40**, D242–D251 (2011).

Distill for CASP10

C. Mirabello[1], G. Tradigo[1,2] , P.Veltri[2], and G. Pollastri[1]

[1] – UCD Dublin, Ireland, [2] – Università di Cosenza, Italy
gianluca.pollastri@ucd.ie

Distill has two main components: a fold recognition stage dependent on sets of protein features predicted by machine learning techniques; an optimisation algorithm that searches the space of protein backbones under the guidance of a potential based on templates found in the first stage. The main differences with our CASP9 systems are: the greatly improved fold recognition stage; the fact that we fit structures directly to the distance maps of templates rather than to predicted contact maps. The difference between Distill and Distill_roll is that for the former we use an improved fold recognition algorithm.

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[1,4] (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 through 3 searches for Distill_roll (PSSM and SAMD profile against PDB sequences and SAMD, with 3 different substitution matrices) and 6 searches for Distill (same as above, plus 3 more searches against PDB PSSM rather than sequences).

In the following stage residue contact maps are predicted by a system based on 2D-Recursive Neural Networks (XXstout[5]). We predict binary maps with a contact threshold of 8Å between Cβ, which are submitted to the RR category. Inputs for map prediction are: the sequence; MSA; PSI-BLAST, SAMD 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 contacts that appear in a weighed average of the distance maps of templates; 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 the template list. For the 5 top-ranked models we reconstruct the backbone with SABBAC, and the full atoms with Scwrl4, then run a brief energy minimisation by gromacs. 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

We await the CASP assessment. On preliminary tests (on the CASP9 set) we have observed a GDT_TS improvement of over 5% over our CASP9 systems.

Availability

http://distillf.ucd.ie/distill/ (Distill), http://dbstill.ucd.ie/distill/ (Distill_roll)

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.

Modeling protein disorder using ESpritz in CASP10

Ian Walsh, Alberto J. Martin Martin, Tomás Di Domenico and Silvio C. E. Tosatto.

Department of Biology, University of Padua, Viale G. Colombo 3, 35131 Padova
ian.walsh@bio.unipd.it

We retrained our server ESpritz[1] on recent PDB structures. The 'E' stands for Efficiency, meaning we designed the algorithm with efficiency as the primary goal. On a single processor it can execute three styles of disorder for entire proteomes in a matter of hours (e.g. human proteome 5-6 hours). The three styles included are X-ray missing atoms, Disprot and NMR based mobility. Disprot based definitions were not executed for this round of CASP since, after benchmarking, X-ray and NMR mobility information were found to correlate better on previous CASP data. NMR mobility is a novel definition derived from our Mobi server[2]. We found that NMR mobility based predictions, executed on NMR targets, increase our performance on CASP data. Benchmarking performed on the now two year old CASP9 data indicated ESpritz would have ranked consistently well with other methods.

Methods

ESpritz is based on amino acid pattern matching algorithms using bidirectional recursive neural networks[3]. ESpritz participated in two groups at this round of CASP. ESpritz group is similar to the method previously published[1]. It is designed to maximize ACC at the expensive if Matthews Correlation coefficient (MCC)[4]. Group ESpritzV2 is a more conservative disorder predictor since it was tuned at a 5% false positive rate. This should improve its MCC at the expensive of the ACC measure. In addition, ESpritzV2 was retrained on more recent structural and sequence information.

ESpritz employs an ensemble based approach where each member of the ensemble contains a different source of input. Sources of input information include: five sequence metrics reflecting different amino acid properties[5], the amino acid itself and evolutionary information in the form of multiple sequence alignments.

In addition to altering the source of input, the disorder definition to be learned is also altered. There are two disorder definitions used at this round of CASP: (1) those with missing backbone C-alpha atoms from X-ray high resolution solved structures and (2) a simple definition based on regions with different conformations among all models in an NMR ensemble[2]. Whenever a CASP target was determined via NMR, predictions were submitted based on NMR tuned algorithms.

Results

All tests were performed on CASP 9 targets, independent from our training sets. Our calculations show ESpritz achieved (sensitivity+specificity)/2=76.9 and an Area Under the receiver operator Curve (AUC) 85.5. ESpritzV2 achieved ACC=(sensitivity+specificity)/2=72.3 and AUC=85.0, however MCC increased by 3 percentage points. These are comparable to the best methods. We anticipate changes in the performance depending on the amount of disorder/mobility present in the NMR structures at this CASP round.

Availability
Both a web server for high-throughput analysis and a Linux executable version of ESpritz are available from: http://protein.bio.unipd.it/espritz/

1. Walsh, I., Martin, A. J. M., Di Domenico, T. & Tosatto, S. C. E. ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics* **28**, 503–509 (2012).
2. Martin, A. J. M., Walsh, I. & Tosatto, S. C. E. MOBI: a web server to define and visualize structural mobility in NMR protein ensembles. *Bioinformatics* **26**, 2916–2917 (2010).
3. Baldi, P., Brunak, S., Frasconi, P., Soda, G. & Pollastri, G. Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics* **15**, 937–946 (1999).
4. Monastyrskyy, B., Fidelis, K., Moult, J., Tramontano, A. & Kryshtafovych, A. Evaluation of disorder predictions in CASP9. *Proteins* **79 Suppl 10**, 107–118 (2011).
5. Atchley, W. R., Zhao, J., Fernandes, A. D. & Drüke, T. Solving the protein sequence metric problem. *PNAS* **102**, 6395–6400 (2005).

Improving remote homologue recognition using evolutionarily conserved topology of protein structures

Chao Wang[1], Bin Ling[1], Qin Huang[1], Haicang Zhang[1], Jin Li[1] and Dongbo Bu[1,*]

[1] – *Institute of Computing Technology, Chinese Academy of Sciences*

FALCON@ict.ac.cn

During the evolutionary process of homologous proteins, some regions are conserved, while some regions are relatively variable, in the sense of both structure and sequence. The conserved regions, denoted as *anchors* in the study, display significant sequence signals, and thus contribute to both alignments and fold recognition, especially for the remote-homologue proteins. Instead of the inefficient multiple structure alignment strategy, an efficient randomized algorithm was proposed to identify structural conserved regions from a set of pairwise structure alignments of templates. Then a target protein was aligned against the structure-conserved anchors rather than the full-length templates. The underlying rational is that alignments might be biased due to the structural variable regions. For the target protein, if likely folds can be recognized with high confidence, the final alignments are built and ranked using HHsearch[1] against likely templates; otherwise, FALCON[2] was executed to construct the full-length structure from the very scratch with distance constraints acquired from alignments against anchors.

Methods

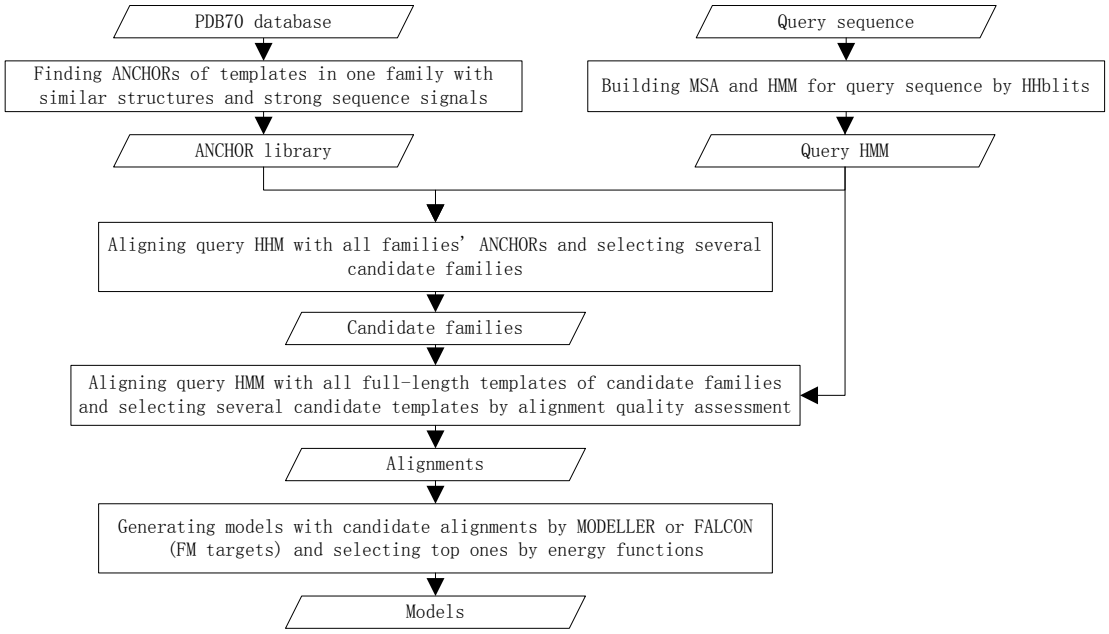


Fig.1 Flowchart of FALCON-TOPO

The flowchart of FALCON-TOPO is depicted in Figure 1. The steps are described in detail as follows:

Step 1. Building ANCHOR database

A database called ANCHOR was built to deposit the regions showing significant structural conservation and strong sequence signal.

More specifically, for each SCOP[3] (version 1.75) family, the evolutionarily conserved regions shared by the homologous proteins in the family were identified as family-specific anchors. For the families that contain only one protein in SCOP, an extension operation was performed to include homologous proteins from PDB70. The criteria of the new proteins to be included are: 1) the new protein should have the same function annotation to the family; 2) the new protein should be structural similar to the existing protein, say RMSD between the two proteins is lower than a threshold.

Given a family of homologous proteins, the structural conserved regions are identified using BLOMAPS[4]. To ensure a strong sequence signal, a further filtering was applied for the corresponding segments reported by BLOMAPS. In particular, the frequencies of amino acids in the corresponding segments are calculated and compared against background amino acid distribution, and the segments with high average K-L distance (say, K-L distance > 2.74) will be filtered out. By this way, these selected anchors are expected to have significant sequence signals.

Step 2. Fold recognition

Given a target protein sequence, we align it against anchors rather than against the full-length templates. The motivation is to avoid the biases rooted in the structural variable regions in templates. Specifically, we design a generative model to describe how a query is generated from a set of anchors of a family. For each family, we calculate the probability that query's profiles are generated from the anchors' profiles, with gap lengths between anchors fitted with a Poisson distribution. The top ones with lowest E-value are kept for the final model generation.

Step 3. Model generating

After recognizing likely folds, the final alignments were generated via aligning query sequence against candidate full-length templates in the family. Briefly speaking, we run HHsearch[4] to build alignments of query against the new template database, and rank alignments by E-value for model generation.

Finally, we generate models by MODELLER for candidate alignments. The generated models are ranked according to dDFIRE[5]energy function. For free-modeling targets, we run FALCON to generate several models and select the best ones by ROSETTA[6] energy function.

1. Söding J (2005). "Protein homology detection by HMM-HMM comparison". *Bioinformatics 21 (7): 951–960.*
2. Shuai Cheng Li, Dongbo Bu, Jinbo Xu, Ming Li, Fragment-HMM: A New Approach To Protein Structure Prediction. *Protein Science, Vol. 17, No. 11, pages 1925-1934, 2008*.
3. Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE. The ASTRAL compendium in 2004. *Nucleic Acids Research 32:D189-D192 (2004).*
4. Wang, S. and Zheng, W. ,Fast multiple alignment of protein structures using conformational letter blocks. *Open Bioinformatics J, volume3,69--83(2009).*
5. Yuedong Yang, Yaoqi Zhou. Specific interactions for ab initio folding of protein terminal regions with secondary structures. *Proteins 2008;72:793-803.*
6. Kim T. Simons, Charles Kooperberg, Enoch Huang and David Baker, Assembly of Protein Tertiary Structures from Fragments with Similar Local Sequences using Simulated Annealing and Bayesian Scoring Functions. *J. Mol. Biol. (1997) 268, 209-225*.

Structure refinement with molecular dynamics simulations in combination with an effective scoring and selection protocol

Vahid Mirjalili[1,3] and Michael Feig[1,2,4]

[1]Departments of Biochemistry and Molecular Biology, [2]Chemistry, [3]Mechanical Engineering, and [4]Computer Science and Engineering , Michigan State University, East Lansing, MI, USA
feig@msu.edu

Molecular dynamics simulations with the new CHARMMM36 force field[1] were carried out followed by model selection and structural averaging.

Methods

Multiple replicas of each target were simulated for 20 ns starting from different random seeds to obtain broad conformational sampling. Explicit solvent simulations were employed, and the systems were neutralized by adding sufficient amount of ions. Cα atoms of residues considered to be accurate in the initial model were weakly restrained to their initial positions, to prevent the structures from drifting away. The set of restraint residues followed CASP suggestions. In cases where there were no suggestions, restraints were applied to core secondary structure elements and in a second set of simulations very weak restraints (k=0.05 kcal/mol/Å^2) were applied to all Cα atoms.

Structural ensembles containing 500 snapshots were created from each simulation. The RMSD from the initial model (iRMSD) and the emprical energy function DFIRE were then used to select a subset of the structures likely to be closest to the native (with low DFIRE scores and low to moderate iRMSD values) using a protocol that was previously optimized based on CASP8 and CASP9 test sets. The selected subset of structures was then used to generate an average structure that was further improved with additional short MD simulations to relieve poor geometries due to the averaging. Additional models submitted to CASP consisted of individual structures with minimal DFIRE scores from the MD ensembles.

Results

Based on preliminary analysis the refinement protocol was able to consistently generate moderately refined structures for almost all targets. Individually selected structures based on DFIRE scores were sometimes refined to a larger extent but often also worse than the initial model thereby lacking overall consistency.

Availability

The protocol combined functionality of the widely disseminated MMTSB tool[2] set in combination with CHARMM and NAMD [3].

1. Best, R.B., et al., (2012) Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain χ1 and χ2 Dihedral Angles. J. Chemical Theory and Comput. **8**, 3257-3273.
2. Feig, M., Karanicolas, J., & Brooks III, C., (2001) MMTSB Toolset, MMTSB NIH Research Resource, The Scripps Research Institute

3. Philips, J.C., et al. (2005) Scalable molecular dynamics with NAMD, J. Comput. Chem. **26**, 1781-1902.

TESTING VARIANTS OF FFAS METHOD IN THE CASP10 EXPERIMENT

Lukasz Jaroszewski, Zhanwen Li, and Adam Godzik

Sanford Burnham Medical Research Institute
Contact: adam@burnham.org

In the CASP010 experiment we tested four variants of the FFAS profile-profile alignment algorithm:

- FFAS03 - the current version of the FFAS algorithm as implemented on the public ffas server at ffas.godziklab.org.
- FFAS03c – in this method a "cascade" of FFAS profile-profile comparison searches is performed for a query sequence. After identifying and aligning the first template structure with the query sequence the covered query fragment is removed. Then new FFAS searches are started for the remaining query fragments and this procedure is continued iteratively until there is no remaining query fragments longer than 30 residues.
- FFAS03jh – same as FFAS03 but sequence profiles are prepared based on JACK-HMMER alignments instead of PSI-BLAST alignments.
- FFAS03mt - same as FFAS03 but multiple templates were used for modeling when available.

Preliminary CASP10 results show that FFASc method preformed better than other variants of FFAS suggesting that the “cascade” searches may be a good approach to improve model’s completeness in cases where the first search yielded only a partial alignment.

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

firestar – ligand binding residue prediction

P. Maietta[1] and M.L. Tress[1]

[1] - CNIO (Spanish National Cancer Research Centre), Madrid, Spain
mtress@cnio.es

Here we describe the protocols used in the prediction of ligand binding sites by the *firestar* server in the 10th edition of CASP. The *firestar*[1,2] server is an expert system for predicting ligand binding and catalytic sites. Predictions are based on the large catalogue of biologically relevant sites culled from PDB structures in the FireDB[3] database. The new version of the *firestar* server requires no human intervention.

Methods

Here we present the new developments of *firestar*[2]. The server extrapolates from a large inventory of functionally important residues, principally from two sources: the biologically relevant small molecule ligand binding residues organized in the FireDB database and the annotated catalytic residues from the Catalytic Site Atlas[4]. *Firestar* makes predictions by homology-based transfer of this functional information. Specifically the server predicts functionally important residues by using local sequence conservation[5].

Several new features have been incorporated into *firestar* to improve the quality of the predictions. All functional residues in the FireDB repository are classified in terms of their biological relevance using evolutionary information, structural data and lists of known cognate ligands.

Previous versions of *firestar* required human interpretation of the results. Now, the whole process has been automatized and a new web interface has been made available. Additionally, the server is able to produce high quality results in a high throughput mode by using sequences as the only input.

The server now also uses HHBlits[6] to generate the initial alignments, thus in theory increasing the coverage.

Results

The *firestar* server returned binding site predictions for 44 targets, sixteen fewer than in CASP9, despite installing HHBlits to increase coverage. Fourteen targets were predicted to bind metals, 27 to bind non-metal biological ligands and five to bind non-cognate ligands (targets could bind more than one type).

Availability

firestar, FireDB and SQUARE can be accessed via http at http://firedb.bioinfo.cnio.es

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-W577;
2. Lopez, G, Maietta, P, Rodriguez, JM, Valencia, A and Tress, ML (2011). *firestar* – advances in the prediction of functionally important residues. *Nucleic Acids Res.* 39

W235-241;
3. 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;
4. Porter, CT, Bartlett, GJ and Thornton, JM (2004). The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res*, 32, D129-D133.
5. Tress, ML, Jones, DT and Valencia, A (2003). Predicting reliable regions in protein alignments from sequence profiles. *J Mol Biol*, 330, 705;

ASTRO-FOLD 2.0: Prediction of Protein Tertiary Structure from First Principles and Global Optimization

G.A. Khoury[1], J. Smadbeck,[1] and C.A. Floudas[1]

[1] – *Department of Chemical and Biological Engineering, Princeton University*
floudas@princeton.edu

ASTRO-FOLD 2.0[1] is a method for the first-principles structure prediction of proteins based on an overall deterministic global optimization framework coupled with mixed-integer optimization (MILP). The novel, six-stage approach combines the classical and new views of protein folding. All methods used were created in-house, and the approach does not use any templates.

Methods

The first stage involves secondary structure prediction using CONCORD.[2] CONCORD is an MILP-based consensus method based on seven secondary structure prediction methods. It combines the strengths of the different methods to maximize the number of correctly predicted amino acids. The second stage focuses on the prediction of the beta-sheet topology using BeST[3] to reduce the three dimensional search space. BeST uses a MILP-based framework to maximize the total strand-to-strand contact potential of a protein. A number of physical constraints to enforce structural and biological validity are applied to provide biologically meaningful topologies. The third stage involves physics-based ILP-driven method[4] for tertiary contact prediction in β, α+β, and α/β proteins. The fourth stage involves the prediction of tight dihedral angle bounds on the via global optimization of loop regions with flexible stems.[5] The fifth stage involves the prediction of the tertiary structure of the full protein sequence. The problem formulation relies on dihedral angle and atomic distance constraints introduced from the previous stages as well as detailed atomistic energy modeling representing a highly nonconvex constrained global optimization problem. The problem is solved using a combination of a deterministically based global optimization approach, αBB[6; 7] and the stochastic global optimization approach Conformational Space Annealing.[8] Once the pool of potential conformers have been generated, ICON,[9] an iterative traveling-salesman problem-based clustering method for identifying near-native protein structures from an ensemble of conformers is applied.

Enhancements to the global optimization portion of the algorithm have been implemented to require an agreement of three different energy functions (ECEPP3[10]/GOAP[11]/AMBER11[12]) before accepting a new locally optimal solution. From initial observations this new consensus requirement substantially improved the bank of conformers as the optimization step is not limited by the accuracy of only one energy function but on the consensus of three different ones.

A method for β-strand structure refinement was developed over the course of the experiment and was applied to the structures produced by ASTRO-FOLD. The method maximized the hydrogen bond network given a beta-sheet topology and subsequently produced final structures with improved secondary structure quality.

Availability

Predictors are welcome to use the individual components of the ASTRO-FOLD 2.0 method via publically available webservers.
CONCORD (2° Structure Prediction): http://helios.princeton.edu/CONCORD/
BeST (Beta-Sheet Topology Prediction): http://selene.princeton.edu/BeST/
ICON (TSP-based clustering): http://helios.princeton.edu/ICON/

1. Subramani, A., Wei, Y. & Floudas, C. A. (2012). ASTRO-FOLD 2.0: An enhanced framework for protein structure prediction. *Aiche Journal* **58**, 1619-1637.
2. Wei, Y., Thompson, J. & Floudas, C. A. (2012). CONCORD: a consensus method for protein secondary structure prediction via mixed integer linear optimization. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science* **468**, 831-850.
3. Subramani, A. & Floudas, C. A. (2012). β-sheet Topology Prediction with High Precision and Recall for β and Mixed α/β Proteins. *PLoS ONE* **7**, e32461.
4. Rajgaria, R., Wei, Y. & 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**, 1825-46.
5. Subramani, A. & Floudas, C. A. (2012). Structure Prediction of Loops with Fixed and Flexible Stems. *The Journal of Physical Chemistry B* **116**, 6670-6682.
6. Klepeis, J. L., Wei, Y., Hecht, M. H. & Floudas, C. A. (2005). Ab initio prediction of the three-dimensional structure of a de novo designed protein: A double-blind case study. *Proteins Structure Function and Bioinformatics* **58**, 560-570.
7. Klepeis, J. L. & Floudas, C. A. (2003). Ab initio tertiary structure prediction of proteins. *Journal of Global Optimization* **25**, 113-140.
8. Lee, J., Scheraga, H. A. & Rackovsky, S. (1997). New optimization method for conformational energy calculations on polypeptides: Conformational space annealing. *Journal of Computational Chemistry* **18**, 1222-1232.
9. Subramani, A., DiMaggio, P. A. & Floudas, C. A. (2009). Selecting High Quality Protein Structures from Diverse Conformational Ensembles. *Biophysical Journal* **97**, 1728-1736.
10. Nemethy, G., Gibson, K. D., Palmer, K. A., Yoon, C. N., Paterlini, G., Zagari, A., Rumsey, S. & Scheraga, H. A. (1992). Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. *The Journal of Physical Chemistry* **96**, 6472-6484.
11. Zhou, H. & Skolnick, J. (2011). GOAP: A Generalized Orientation-Dependent, All-Atom Statistical Potential for Protein Structure Prediction. *Biophysical Journal* **101**, 2043-2052.
12. Weiner, P. K. & Kollman, P. A. (1981). AMBER: Assisted model building with energy refinement. A general program for modeling molecules and their interactions. *Journal of Computational Chemistry* **2**, 287-303.

Hydrogen Bond Network Optimization for Improved Tertiary Structure Refinement

J. Smadbeck[1], G.A. Khoury[1] and C.A. Floudas[1]

1 – Department of Chemical and Biological Engineering, Princeton University, Princeton, NJ 08540
jsmadbec@princeton.edu

Recent studies have shown that the hydrogen bond count in proteins has a strong, linear correlation to the molecular weight of the protein[1]. However, many Tertiary Structure Prediction methods produce structures with hydrogen bond counts well below the number expected. We have developed a method to optimize the number of hydrogen bonds of a given structure, while minimizing the necessary structural changes to do so. The method produces refined structures with well defined secondary structure, even when the starting structure has a little discernible secondary structure.

Methods

The refinement method takes in a starting predicted tertiary structure of a protein as input. Both atom-atom distance constraints and dihedral constraints are calculated for the entire structure. These constraints are necessary to limit the overall structural changes induced by the refinement method. The ideal hydrogen bond network is calculated from structural analysis of the given, starting protein. User input of known secondary structure or predicted secondary structure and topology (CONCORD[2]/BeST[3]) is used to improve the hydrogen bond network optimization performance. Constraints were also derived for the total number of hydrogen bonds based on the molecular weight of the input protein[1].

The derived atom-atom distance constraints are used as input to the CYANA 2.1 software package for NMR structure refinement[4]. CYANA is used to produce a number of candidate structures that satisfy the derived hydrogen bond network, with minimal violation of the starting distance and angle constraints. The combination of the two constraint sets limits the overall structural changes allowed during a single iteration of the method, while producing structures with improved secondary structure and hydrogen bonding.

The structure with the best chance at improving GDT value to the native is chosen from the large ensemble of produced structures through energetic (GOAP[5]/DFIRE[6]) and structural (Hydrogen Bond number, Solvent Accessible Surface Area, etc.) analysis. Testing was also done with the aim of using the hydrogen bonding network optimization derived structures as input to a well established refinement method, such as Kobamin[7-9]. An example of a structure produced by the Floudas group's ASTROFOLD 2.0[10-13] method is shown in Figure 1, highlighting the secondary structure improvements that the method is capable of through hydrogen bond network optimization.

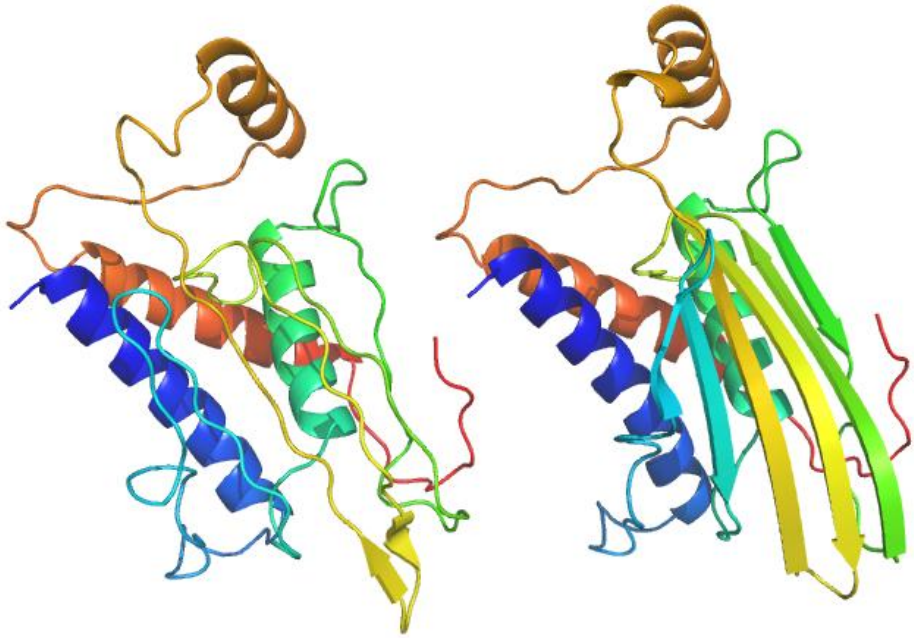


Figure 1: CASP 10 Target T0676 before (left) and after (right) hydrogen bond network optimization. Starting structure output from ASTROFOLD 2.0 structure prediction.

1. Glyakina A.V., Bogatyreva N.S., Galzitskaya O.V. (2011). Accessible surfaces of beta proteins increase with increasing protein molecular mass more rapidly than those of other proteins. PLoS One 6: e28464.
2. Wei Y., Thompson J., Floudas C.A. (2012). CONCORD: a consensus method for protein secondary structure prediction via mixed integer linear optimization. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science 468, 831-850.
3. Subramani A., Floudas C.A. (2012). β-sheet Topology Prediction with High Precision and Recall for β and Mixed α/β Proteins. PLoS ONE 7, e32461.
4. Guntert P. (2004) Automated NMR structure calculation with CYANA. Meth. Mol. Biol. J. Mol. Biol., 278 (2004), pp. 353–378
5. Zhou, H., Skolnick, J. (2011). GOAP: A Generalized Orientation-Dependent, All-Atom Statistical Potential for Protein Structure Prediction. Biophysical Journal 101, 2043-2052.
6. Yang Y., 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,17:1212-1219.
7. Chopra G., Summa C.M., Levitt M.(2008). Solvent dramatically affects protein structure refinement. Proc Natl Acad Sci USA (2008) vol. 105 (51) pp. 20239-44
8. Chopra G., Kalisman N., Levitt M.(2010). Consistent refinement of submitted models at CASP using a knowledge-based potential. Proteins (2010) vol. 78 (12) pp. 2668-78
9. Rodrigues J.P.G.L.M., Levitt M., Chopra G. (2012). KoBaMIN: a knowledge-based minimization web server for protein structure refinement *(2012) 40(W1): W323-W328*
10. Subramani A., Wei Y.,Floudas C.A. (2012). ASTRO-FOLD 2.0: An enhanced framework for protein structure prediction. Aiche Journal 58, 1619-1637.
11. Klepeis J.L., Floudas C.A. (2002) ASTRO-FOLD: Ab-initio secondary and tertiary structure prediction in protein folding, Computer Aided Chemical Engineering, Volume 10, Pages 97-102

12. Klepeis J.L., Wei Y., Hecht M.H., Floudas C.A. (2005). Ab initio prediction of the three-dimensional structure of a de novo designed protein: A double-blind case study. Proteins Structure Function and Bioinformatics 58, 560-570.
13. Klepeis, J.L., Floudas, C.A. (2003). Ab initio tertiary structure prediction of proteins. Journal of Global Optimization 25, 113-140.

Multiplayer online game-based homology and ab-initio modeling

F. Khatib[1], J. Flatten[1], T. Husain[1], K. Xu[1], S. Cooper[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. CASP10 targets shorter than 170 residues 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 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 CASP10 puzzle was typically accessible to Foldit players for 8-9 days.

For CASP10 targets shorter than 170 residues in the "All Groups" category, two different Foldit puzzles were given to the players. One puzzle started from an extended chain, with alignments to known templates taken from the RaptorX[3], Sparks-X[4], and HHsearch[5] servers provided. 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 in-game tools listed above. For the second puzzle, models were constructed using the Zhang-Server[6] and QUARK[7] predictions. These server models were initially minimized using Rosetta and then given as starting points for the Foldit players to refine. This same protocol was used for CASP10 targets in the "Refinement" category, where server models were first minimized with Rosetta before being given to the Foldit players.

For each CASP10 puzzle, Foldit players were provided with secondary structure predictions generated by the SAM-T08 server[8], in the form of a sequence logo. For some of the CASP10 puzzles, the top scoring Foldit model was not submitted by our group, but rather by the wfFUIK CASP10 group. In the collaborative spirit of the WeFold project, we did not want to submit redundant predictions. Therefore, a complete assessment of all Foldit results should also include these submissions.

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 http://tinyurl.com/academic-foldit .

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. *Methods in enzymology* **487**, 545-74.

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. Peng,J. & Xu,J. (2009) Boosting Protein Threading Accuracy. *Research in Computational Molecular Biology (RECOMB)*, **5541**, 31-45.
4. Yang,Y., Faraggi,E., Zhao, H. & Zhou,Y. (2011) Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of the query and corresponding native properties of templates. *Bioinformatics* **27**, 2076-2082.
5. Söding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics.* **21**(7):951-60.
6. Zhang,Y. (2008) I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*, vol 9, 40.
7. Xu,D. & Zhang,Y. (2012) Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins* **80**, 1715-35.
8. Karplus,K. (2009) SAM-T08: HMM-based Protein Structure Prediction. *Nucleic Acids Research.* **37**(2): W492-7.

Four Body Potentials

Eshel Faraggi, Andrzej;Kloczkowski

Battelle Center for Mathematical Medicine, NCH

This approach is a repetition of the four body potential approach of the previous CASP. No changes were made to the program except some minor code revisions to enable compilation on local machines. This server will eventually be part of the bioinformatics services page of the Kloczkowski Lab at the Battelle center.

Structure refinement using energy-guided fragment regrowth

Samuel W.K. Wong[1], Valeria Espinosa[1], Kevin Bartz[1], S.C. Kou[1], Jun S. Liu[1] and Jinfeng 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

Refinement of template structures continues to be a challenging problem in protein structure prediction. In the CASP9 refinement experiment[1], attempts to refine the best server model often did not yield substantial improvement in backbone geometry and other evaluation metrics. Generally, failure to improve a template structure can be attributed to a combination of sampling inefficiencies and errors in the scoring function. We designed an updated version of FRESS (fragment re-growth via energy-guided sequential sampling) to sample candidate conformations, and developed GAMs (generalized additive models) to select the final predicted structure from the sampled conformations.

Methods

Input: For regular CASP10 TS targets, we first used an automated HHsearch[2] script to detect homologous proteins, followed by a Modeller[3] run to build a selection of initial models based on the matching templates. For refinement targets, we began with the provided template model. Reduce[4] was run on all models for adding hydrogen atoms to the structures.

Sampling: The basic Monte Carlo move in our simulations is based on FRESS, which we originally developed in the context of HP models[5]. Applied to protein structures, our method is an efficient way to sample the conformation of a fragment while keeping the rest of the structure fixed. In each chosen fragment, residue backbone atoms are regrown one at a time, with the goal of finding alternative closed and sterically feasible conformations with low energy. The sampled torsion angles at each step are biased towards the eventual closure of the fragment, and conditioned on the secondary structure type of the fragment being considered. Fragments are then completed via the addition of side chains to feasible backbones, and evaluated using the full energy function. The energy function at this step employs a weighted combination of Van der Waals energy, along with statistics-based terms DFIRE2[6], Oscar-o[7], and in-house implementations of hydrogen bonding and backbone torsion terms. To improve the ability of the simulations to move through local energy barriers, we implemented PTEEM[8] (parallel tempering with equi-energy moves) to govern the global parallelization scheme over our computing cluster.

Selection: During the sampling phase, conformations are saved at set intervals, creating a pool of structures from which the final prediction must be selected. The prediction could be selected by simply choosing the lowest energy structure; however our energy function is not sufficiently accurate. To improve the final prediction, we built a selection model based on GAM[9]. The terms used to build the model include the energy terms listed above, and additionally a secondary structure score based on differences from the PsiPred[10] prediction for the sequence, and the OPLS-AA[11] force field. Structures used for building this statistical model were taken from the CASP9 experiment. The best GAM-scores on predicted GDT-TS were chosen as the prediction.

Results

Using a cross-validation approach, the method was tested on CASP9 refinement targets, obtaining a mean improvement of 0.011 GDT-TS units over the starting model.

Availability

The executable implementing the FRESS method is written in C++ and available upon request.

1. MacCallum, J.L., Perez, A., Schneiders, M.J., Hua, L, Jacobson, M.P., & Dill, K.A. (2011). Assessment of protein structure refinement in CASP9. *Proteins* **79**(Suppl 10), 74-90.
2. Soding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**, 951-960.
3. N. Eswar, M. A. Marti-Renom, B. Webb, M. S. Madhusudhan, D. Eramian, M. Shen, U. Pieper, A. Sali (2006). Comparative Protein Structure Modeling With MODELLER. Current Protocols in Bioinformatics, Supplement 15, 5.6.1-5.6.30.
4. Word, J.M., Lovell, S.C., Richardson, J.S., Richardson, D.C. Asparagine and Glutamine: Using Hydrogen Atom Contacts in the Choice of Side-chain Amide Orientation. *J. Mol. Biol.* **285**, 1735-1747.
5. Zhang, J., Kou, S.C., Liu, J.S. (2007). Biopolymer structure simulation and optimization via fragment regrowth Monte Carlo. *J. Chem. Phys.*, **126**, 225101.
6. Yang, Y., & 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 Sci.* **17**, 1212-1219.
7. Liang, S., Zhou, Y., Grishin, N., & Standley, D.M. (2011). Protein side chain modeling with orientation-dependent atomic force fields derived by series expansions. *J. Comput. Chem.* **32**, 1680-1686.
8. Baragatti, M., Grimaud, A., & Pommeret, D. (2011). Parallel Tempering with Equi-Energy Moves. Working paper, retrieved from http://arxiv.org/pdf/1101.4743.pdf
9. Wood, S.N. (2004) Stable and efficient multiple smoothing parameter estimation for generalized additive models. *J. Amer. Statist. Ass.* **99**, 673-686.
10. Jones, D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol*. **292**, 195-202.
11. 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.

GOAPQA server for quality assessment prediction in CASP10

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 GOAPQA server is an implementation of the GOAP[1] statistical potential for protein model quality assessment prediction.

Method

The new statistical potential, GOAP (Generalized Orientation-dependent All-atom Potential), depends on the relative orientation of the planes associated with each heavy atom in interacting pairs. GOAP is a generalization of previous orientation-dependent potentials that consider only representative atoms or blocks of side-chain or polar atoms. GOAP is decomposed into distance and angle-dependent contributions. The DFIRE[2] distance–scaled finite ideal gas reference state is employed for the distance-dependent component of GOAP. GOAP was tested on eleven commonly used decoy sets containing 278 targets, and recognized 226 native structures as best from the decoys, whereas DFIRE recognized 127 targets. The major improvement comes from decoy sets that have homology modeled structures that are close to native (all within ~ 4.0 Å or from the ROSETTA ab initio decoy set. For these two kinds of decoys, orientation independent DFIRE or only side-chain orientation-dependent RWplus[3] performed poorly. While the OPUS-PSP[4] block-based orientation-dependent, side-chain atom contact potential performs much better (recognizing 196 targets) than DFIRE, RWplus and dDFIRE[5] it is still ~15% worse than GOAP. Encouraged by its performance in benchmarking, we applied GOAP to protein model quality assessment prediction. In order to convert the raw GOAP score into model quality score between 0 and 1, for each model, its average TM-score[6] to the top five GOAP ranked models is used as its quality score. Thus GOAPQA in its current simple form works only for assessing more than five models.

Result

For the released 41 targets (up to Sep. 20, 2012, excluding cancelled targets, no domain parsing), we compare the performance of GOAPQA to the 3D-jury[7] procedure using the sum of a model's TM-score to all other models for ranking (the larger the sum, the better the model). In Table 1, we show that for stage 1 prediction, the average correlation of predicted quality to real model quality of 3D-jury is better than that of GOAPQA whereas for stage 2, GOAPQA is better. This demonstrates that a non-consensus method as GOAPQA can perform better than consensus-based 3D-jury and the released models at stage 1 are not challenging enough for naïve consensus-based approaches.

Table 1
Average correlations of predicted and real model quality as measured by GDT-TS-score.

Stage 1		Stage 2	
GOAPQA	3D-jury	GOAPQA	3D-jury
0.58	0.71	0.50	0.46

Availability

The GOAP program is available at http://cssb.biology.gatech.edu/

1. Zhou, H. & Skolnick, J. 2011. GOAP: A Generalized Orientation-Dependent, All-Atom Statistical Potential for Protein Structure Prediction. Biophysical Journal **101**, 2043-2052.
2. Zhou, H., and Y. Zhou. 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.
3. Zhang, J., and Y. Zhang. 2010. A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. Plos ONE 5:e15386.
4. Lu, M., A. Dousis, and J. Ma. 2008. OPUS-PSP: An orientation-dependent statistical all-atom potential derived from side-chain packing. J. Mol. Biol. 376:288-301.
5. Yang, Y., and Y. Zhou. 2008. Specific interactions for ab initio folding of protein terminal regions with secondary structures. Proteins 72:793-803.
6. Zhang, Y., and J. Skolnick. 2004. A scoring function for the automated assessment of protein structure template quality. Proteins 57:702--710.
7. Ginalski K, Elofsson A, Fischer D, Rychlewski L. 3D-jury: a simple approach to improve protein structure predictions. Bioinformatics 2003;19:1015-1018.

GOBA: Gene Ontology-Based Assessment of protein structural models.

B.M. Konopka[1], M. Kurczynska[1], P. Gasior[1], T. Golda[1], P. Wozniak[1], J-C. Nebel[2], M. Kotulska[1]

[1] – *Institute of Biomedical Engineering, Wroclaw University of Technology, Poland,*

[2] - *Faculty of Science, Engineering and Computing, Kingston University, Kingston-upon-Thames, United Kingdom*

bogumil.konopka@pwr.wroc.pl

GOBA[1]is a quasi-single model quality assessment program. It estimates the compatibility between the structure of a protein model and its expected function. The approach is based on the assumption that a high quality model is expected to be structurally similar to proteins functionally similar to the prediction target. Whereas DALI[2]is used to measure structure similarity, protein functional similarity is quantified using standardized and hierarchical description of proteins provided by Gene Ontology[3]combined with Wang's algorithm[4]for calculating semantic similarity. During CASP10 our main focus was to submit predictions in QA stage1 and stage2 categories (gr # 031 & 033), however we also submitted predictions in human TS category(gr# 131).

Methods

GOBA is based on the protein structure-function relation therefore a perquisite for using GOBA is the availability of Molecular Function GO terms annotated to the target protein. These can be for instance taken from the target Uniprot record. During the contest we used AmiGO[5] tool (http://amigo.geneontology.org/cgi-bin/amigo/blast.cgi) and an E-value threshold of $5.6*10^{-6}$ for assinging GO terms - we used human curated annotations only (nonIEA).

The quality assessment procedure is the following. The target functional annotations and the model-structure are fed into GOBA. The algorithm starts by constructing a Similarity List (Slist) that is a list of Structural Neighbors (SNs) of the model. SNs are identified using DALI. This is followed by calculations of the Functional Similarity (FS) score between every SN and the target protein. Wang's semantic similarity algorithm is used for FS calculation.

Once the Slist is produced, model-structure quality scores are calculated using variations of the Receiver Operating Characteristic (ROC) methodology. We proposed two diverse approaches to plotting the ROC curve. In both approaches, the SNs are divided into two sets. If the FS of a Structural Neighbor is greater than a set FS threshold parameter, it is classified as a positive hit, otherwise it is a negative hit.

In the basic procedure, a sensitivity vs specificity ROC curve is plotted for the Slist, with the DALI Z-score assumed as the cut-off parameter. The basic model quality score – the GA-score - is defined as the Area Under the plotted ROC Curve. In the modified approach – yGA-score - the DALI Z-score of each Structural Neighbor is explicitly used when ploting the ROC and calculating the AUC.

Both GA and yGA scores depend on the FS threshold as a parameter. In the contest we tested linear combinations of scores calculated for three FS thresholds, i.e. 0.5, 0.7 and 0.9. The

submitted quality score of a model was an average of GA scores (gr# 031) or yGA scores(gr# 033) calculated for listed thresholds.

GA scores are “single model” MQAPs since they provide an absolute model-structure quality
score only based on a single model. On the other hand, the yGA scores rely on Dali Z-score values which are not absolute measures of structural similarity since Z-score values are only defined within a given population of structures. Consequently, yGA based GOBA can only be used for comparing model-structures of a given target. However, contrary to consensus (clustering-based) methods (which require large sets of structures) yGA scores can be used with model sets of any size.

The BioNanopore group (gr# 131) used GOBA in a simple structure prediction pipeline. A number of prediction servers available on-line were used to generate a pool of putative model-structures. Models were ranked using the yGA_579 score (described above) and top five structures were submitted to the contest. The servers used are listed in Table 1. The group processed 'All groups' category models only.

Table 1: List of servers used in TS category by group #131.

Pcons.net	http://pcons.net	MULTICOM	http://casp.rnet.missouri.edu/multicom_3d.html
SAM-T08	http://compbio.soe.ucsc.edu/SAM_T08/T08-query.html	Phyre	http://www.sbg.bio.ic.ac.uk/~phyre/
(PS)2-v2	http://ps2v2.life.nctu.edu.tw/	SwissModel	http://swissmodel.expasy.org/workspace/index.php?func=modelling_simple1
AS2TS	http://proteinmodel.org/AS2TS/AS2TS_MB/index.html	(PS)2	http://ps2.life.nctu.edu.tw/
LOOPP	http://clsb.ices.utexas.edu/loopp/web/	3D-JIGSAW	http://bmm.cancerresearchuk.org/~populus/populus_submit.html
SPARKSX	http://sparks.informatics.iupui.edu/yueyang/sparks-x/	ESyPred3D	http://www.fundp.ac.be/sciences/biologie/urbm/bioinfo/esypred/

Results

The main limitation of the method is the availability of GO term annotations. After removing canceled targets, 102 CASP10 targets were issued in 'All groups' (47) and 'Server Only' (55) TS categories. Despite a strict E-value criterion, we managed to acquire good annotations for 45 of them. This shows that the method is applicable to a significant number of cases.

Availability

GOBA quality assessment program is available for download at:
http://www.ibp.pwr.wroc.pl/KotulskaLab/materialy/GOBA%20-%20Model%20Quality%20Assessment%20Programe/GOBA_src_BMC_BIO.tgz
It requires DALI installed locally.

1. Konopka BM., Nebel JC, Kotulska M (2012), Quality assessment of protein model-

structures based on structural and functional similarities. *BMC Bioinformatics* **13**:242
2. Holm L, Rosenström P (2010),Dali server: conservation mapping in 3D. *Nucleic Acids Res.* **38**:W545–W549.
3. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000). Gene ontology: tool for the unification of biology. *Nat Genet*, **25**(1):25–29.
4. Wang JZ, Zhidian D, Rapeeporn P, Yu PS, Chin-Fu C (2007). A new method to measure the semantic similarity of GO terms. *Bioinformatics*, **23** :1274–1281. :10.1093/bioinformatics/btm087.
5. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S, AmiGO Hub, Web Presence Working Group (2009). AmiGO: online access to ontology and annotation data. . Bioinformatics, **25**(2):288-9.

G-QA: Group based Quality Assessment of protein

Juexin Wang[1,2], Yanchun Liang[1]

[1] *Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, College of Computer Science and Technology, Jilin University, Changchun , China,* [2] *Department of Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw Avenue, Ann Arbor, MI, USA*
juexinwang@gmail.com, liangyc@jlu.edu

Introduction

The model quality assessment of protein (MQAP) is an essential challenging issue in protein structure prediction. Among all MQAP methods in QA category of last several CASP competitions, consensus methods incorporating all the model information together in calculating similarities could obtain much more accurate results. In QA category of CASP10, only a few models and good part of all models are provided to QA participants in stage 1 and stage 2, which eliminates greatly the power of existing consensus methods in QA.

Methods

In aim to use the great power of consensus methods, we developed Group based Quality Assessment of Protein (G-QA) method. The basic idea of G-QA is to use as many reference models as possible to facilitate the usage of information gathering by consensus methods. We used referenced models generated by I-TASSER[1] Monte Carlo Simulation in supplementing the models provided by CASP organizer, and then use modified consensus methods on these enhanced model pools in calculating the paired tertiary structure similarities as the QA scores.

Availability

G-QA is NOT available for public now.

1. Yang Zhang (2008). I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*, 9, 40.

Julia Handl, Shaun Kandathil, Simon Lovell

University of Manchester

A core interest in our group is to better understand and improve fragment-assembly approaches to ab initio protein structure prediction.

Low-resolution methods for protein structure prediction continue to play a crucical role in ab initio prediction, as they are essential for providing suitable starting points for subsequent full-atom refinement. Recent CASP experiments show that existing methods for low-resolution sampling have a tendency to break down for proteins with > 70 amino acids, and the analysis of the corresponding search trajectories reveals that this is, at least in part, due to a break-down of the sampling protocols employed.

An improved understanding of the properties and search dynamics of fragment assembly approaches is important to enable the development of more effective sampling protocols that will scale to larger proteins. For this reason, we have recently conducted a systematic analysis of the role of fragments during the search process: our analysis in [1] highlights the dual role that fragments play during low-resolution sampling, as they constrain the search space but simultaneously define the size of the variation operator. This suggests that it may be valuable to isolate these two different aspects of fragments, which we have termed as "fragment length" and "move length", and this idea was further explored in [1].

During this CASP competition, we utilized a modified version of Rosetta that separates these concepts of fragment and move length. The emphasis was on exploring the use of medium to large fragments in combination with shorter move sizes. Using this method, ab initio predictions were made for the majority of CASP targets (where time constraints permitted), and different combinations of fragment length / move length were utilized to generate models of different rank. The computational resources utilized were fairly small with 500 low-resolution models generated per target only.

1. J. Handl, J. Knowles, R. Vernon, D. Baker and S.C. Lovell, The dual role of fragments in fragment-assembly methods for de novo protein structure prediction, Proteins 80(2):490-504

Server-based fold recognition predictions using hGen3D & NewServ

D.W.A. 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/psipred

The UCL NewSerf and hGen3D servers implement a fully automated template selection and homology modelling strategy.

Methods
NewSerf is an update of our automated homology pipeline, BioSerf, which was entered in to CASP8 and CASP9. This method has been substantially streamlined and all *de novo* modelling components have been removed. The NewSerf process initially attempts to find appropriate templates for homology modelling using PSI-BLAST [1] against the fasta PDB, PSIPRED [1] & pGenTHREADER [2] and HHBlits[4]. Possible templates are then selected with conservative scoring cutoffs; E-values less than $5x10^{-5}$ or $1x10^{-3}$ for PSI-BLAST and HHBlits respectively or a GenTHREADER p-value of <= 0.01. Next we attempt to 'intelligently' select appropriate homology modelling templates from the set of all good scoring putative templates. First QMODCHECK [5] is run to select the compatibility of the modelling target sequence with each of the putative templates. Any templates that do not score well are discarded. Then we use a TM score Jury method to select a maximum of 10 templates. All-against-all TM scores[6] are calculated for the remaining set of templates, any putative template that does not cluster with the majority of the templates is discarded and then the 10 most tightly clustered templates are selected for homology modelling. Lastly MODELLER [7] is used, with the alignments generated by PSI-BLAST, pGenTHREADER and HHBlits, to generate a model using the 10 template structures that were chosen in the TM Jury step.

hGen3D is an experimental modification of NewSerf which attempts to select the optimal number of templates by building models starting with the top-ranked template and ending with a maximum of 10 templates. From these (maximum of 10) models the best model is selected using two model quality assessment programs (QMODCHECK and MODELLER'S DOPE score).

Results
Models for all server targets were submitted.

Availability
NewSerf (BioSerf2) can be access from the following URL: http://bioinf.cs.ucl.ac.uk/psipredtest

1. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997 Sep 1;25(17):3389-402.
2. Jones, D.T. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol (1999), 292, 195–202.
3. 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.
4. Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods. 2011 Dec 25;9(2):173-5. doi: 10.1038/nmeth.1818.
5. Pettitt CS, McGuffin LJ, Jones DT. Improving sequence-based fold recognition by using 3D model quality assessment. Bioinformatics. 2005 Sep 1;21(17):3509-15. Epub 2005 Jun 14.
6. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score.
7. Nucleic Acids Res. 2005 Apr 22;33(7):2302-9. Print 2005.
8. Eswar, N., Eramian, D., Webb, B., Shen, M.Y. & Sali, A. Protein structure modeling with MODELLER. Methods Mol Biol (2008), 426, 145–159.

Template-based structure prediction with HHpredA

M. Meier[1], A. Meier[1], C. Angermüller[1] and J. Söding[1]

[1] *Gene Center Munich, LMU*
meiermark|meier|angermueller|soeding@genzentrum.lmu.de

We registered three servers for CASP10: HHpredA is almost identical to HHpredA from CASP9. HHpredA-thread includes profile-to-structure scoring. HHpredB was intended to perform model quality estimation via a newly implemented method but did not get ready in time and was identical to HHpredA. As in CASP9, we tried to keep response times under 10 min.

Methods
We summarize the HHpredA pipeline in the following:

1. The *first* template was selected by a regression neural network that predicts the model quality (TM-score) based on four input features from the HHsearch results: HHsearch raw score, secondary structure score, template resolution, and length-normalized sum of posterior probabilities over all aligned residues.
2. Further 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 the 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 replaced MODELLER's[4] distance constraints to account for the varying confidence of aligned residue pairs along the alignment, again measured by the posterior probabilities. We define bimodal distance restraints 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, a neural network designed for training the parameters of a mixture of Gaussians. Badly aligned residues with low posteriors will lead an increased background mixture weight and larger sigmas of the Gaussian mixture components.
4. We performed three search iterations with our program HHblits[3] with default parameters through the uniprot20 database of HMMs to build a multiple sequence alignment (MSA) for the query sequence. The query alignment was converted into an HMM with HHmake, and HHsearch[2] from the HH-suite[3] was used to search for templates in representative HMMs of the PDB (70% maximum sequence identity). These representative template HMMs were also generated using three iterations of HHblits with default parameters.

Availability
HHpred, HHblits and more bioinformatics tools are available at our bioinformatics toolkit[5] at *http://toolkit.lmb.uni-muenchen.de/*.

References
1. Hildebrand A, Remmert M, Biegert A, and Söding J. (2009) Fast and accurate automatic structure prediction with HHpred. *Proteins* **77** Suppl 9: 128-132.

2. Söding J. (2003) Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**: 951-960.
3. Remmert M, Biegert A, Hauser A, and Söding J. (2011) HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9**: 173-175.
4. Sali A, Blundell TL. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol. Biol.* **234**: 779-815.
5. Biegert A, Mayer C, Remmert M, Söding J, Lupas A N. (2006) The MPI Bioinformatics Toolkit for protein sequence analysis. *Nucleic Acids Res.* **34**: W335-339.

Prediction of functional sites with HHpredA-func

M. Meier[1], A. Meier[1], C. Angermüller[1] and J. Söding[1]

[1] *Gene Center Munich, LMU*
meiermark|meier|angermueller|soeding@genzentrum.lmu.de

HHpredA included a functional site prediction module **HHpredA-func**. It searches for homologous templates with annotated functional sites in the fireDB[1] and assesses the reliability of these predicted functional sites based on match probability, query-template alignment accuracy, conservation of residues within the functional site between query and template, and conservation of these residues within the query multiple sequence alignment. If the template-based prediction confidence is below 0.3 for the best predicted site, the de-novo prediction method Frpred[2] is employed instead.

Availability
HHpred, HHblits and more bioinformatics tools are available at our bioinformatics toolkit[3] at *http://toolkit.lmb.uni-muenchen.de/*.

1. Lopez G, Valencia A, and Tress M. (2007) FireDB -- a database of functionally important residues from proteins of known structure. *Nucleic Acids Res.* **35**: D219—D223.
2. Fischer J, Mayer CE, and Söding J. (2008) Prediction of protein functional residues from sequence by probability density estimation. *Bioinformatics* **24**: 613-620.
3. Biegert A, Mayer C, Remmert M, Söding J, Lupas A N. (2006) The MPI Bioinformatics Toolkit for protein sequence analysis. *Nucleic Acids Res.* **34**: W335-339.

Extending HHpred by sequence-structure threading: HHpredA-thread

M. Meier[1], A. Meier[1], C. Angermüller[1] and J. Söding[1]

[1] *Gene Center Munich, LMU*
meiermark|meier|angermueller|soeding@genzentrum.lmu.de

Methods

HHpredA-thread is similar to HHpred but, in addition to the amino acid column score and the secondary structure similarity score, we developed three statistical scores that compare the query's sequence profile with the template structure and template profile in order to improve the ranking of templates and the query-template alignments. First, both the query and template profiles are discretized by assigning each profile column the most similar *column state* out of a previously learned alphabet of size 62. We also count the number of residue contacts for each residue in the template and in the virtual query model that the query-template alignment would give rise to. Using these data, number of residue-residue contacts and column states for query and template, we computed the following residue-wise log-odds scores:

The column state substitution score evaluates column state substitutions given the number of contacts in the template. The compactness score assesses the conservation of the number of contacts between the model and the template given the template's column state. The model quality score evaluates the likelihood ratio of all pairwise distances in the virtual query model given the column states of query and template at the alignment position. The three threading scores are linearly combined with optimized weights to obtain the total threading score function.

Availability

HHpred, HHblits and more bioinformatics tools are available at our bioinformatics toolkit[1] at *http://toolkit.lmb.uni-muenchen.de/*.

1. Biegert A, Mayer C, Remmert M, Söding J, Lupas A N. (2006) The MPI Bioinformatics Toolkit for protein sequence analysis. *Nucleic Acids Res.* **34**: W335-339.

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

J. Bacardit, P.Widera and N. Krasnogor

School of Computer Science, University of Nottingham
jaume.bacardit@nottingham.ac.uk

Our CASP10 residue-residue contact prediction method is a small variation of the method which has participated in CASP8 and CAPS9[1]. The predictions are performed by an ensemble of 1250 rule sets that are generated by BioHEL[2], our in-house machine learning system. Three types of input information were used to train our system: (1) detailed local sequence information from two selected regions (windows) around specific residues, (2) information about the connecting segment between the two target residues and (3) global sequence information.

Two windows of ±4 residues were constructed around the two target residues. Each residue in the windows was characterised using: (1) a position-specific scoring matrix (PSSM) profile computed with PSI-BLAST[5], (2) secondary structure (SS) predicted by PSIPRED[7], (3) five-state coordination number (CN)[8], (4) five-state relative solvent accessibility (SA)[3] and (5) five-state Recursive Convex Hull (RCH)[3]. CN, SA and RCH were predicted using BioHEL as well.

The connecting segment was represented by the distributions of amino acids types, predicted secondary structure states[6], as well as predicted CN, SA and RCH. The global sequence information contained 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[6] and the contact propensity between the amino acid types of the target residue pair[9]. In total, 511 variables were used in the training process.

The training process followed the four steps below:

1. We selected a set of 3262 protein chains from PDB-REPRDB with a resolution less than 2Å, less than 30% sequence identity and without chain breaks or 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 720000 residue pairs with a fixed 2:1 proportion of non-contacts to real contacts.
4. We run BioHEL 25 times for each training sample with different initial random seeds, thus generating an ensemble of 1250 rule sets (50 training samples x 25 seeds). This ensemble performed the residue-residue contact prediction.

The changes between this and our CASP9 predictor are that (1) we have simplified our representation and (2) we have increased the sizes of the samples. We have removed from our previous representation a third window of residues placed at the middle point between the target pair of residues. Our analysis of the rule sets involved in the previous representation[1] suggested

that this third window was the weakest contributor to the prediction capacity of our method. Also, we have increased the sizes of the samples from 660000 to 720000 residue pairs.

Availability

Our contact map prediction method is available as part of the *ICOS server for the prediction of structural aspects of protein residues*, at http://icos.cs.nott.ac.uk/servers/psp.html

1. J. Bacardit, P. Widera, A. Marquez-Chamorro, F. Divina, J.S. Aguilar-Ruiz and Natalio Krasnogor. Contact map prediction using a large-scale ensemble of rule sets and the fusion of multiple predicted structural features. Bioinformatics. First published online July 25, 2012 doi:10.1093/bioinformatics/bts472
2. J. Bacardit, E.K. Burke and N. Krasnogor. Improving the scalability of rule-based evolutionary learning. Memetic Computing journal 1(1):55-67, 2009
3. 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.
4. 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.
5. 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.
6. M. Punta and B. Rost (2005) "Profcon: novel prediction of long-range contacts". Bioinformatics 21(13):2960-8.
7. D.T. Jones, (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* ***292***, 195-202.
8. J. Bacardit, M. Stout, J.D. Hirst, A. Valencia, R.E. Smith and N. Krasnogor. Automated Alphabet Reduction for Protein Datasets. BMC Bioinformatics 10:6, 2009
9. G. Shackelford and K. Karplus. (2007) Contact Prediction using Mutual Information and Neural Nets. *Proteins: Structure, Function, and Bioinformatics*, 69(S8):159-164.

CMAPpro: Deep Architecture for Contact Map prediction

P Di Lena, K. Nagata and P. Baldi

Department of Computer Sciente, Institute for Genomic and Bioinformatics
University of California, Irvine
pdilena@uci.edu, knagata@uci.edu, pfbaldi@ics.uci.edu

CMAPpro[3] is a Deep Neural Network architecture for residue-residue contact prediction. The deep architecture is designed to address two main issues in contact prediction: (1) Residue-residue contacts are not randomly distributed in native protein structures, rather they are spatially correlated. Contact predictors generally do not take into account these correlations, not even at the local level, since the contact probability for a residue pair is typically learned/inferred independently of the contact probabilities in the neighborhood of the pair. (2) Proteins do not assume a 3D conformation instantaneously, but rather through a dynamic folding process that progressively refines the structure. In contrast, machine learning approaches typically attempt to learn contact probabilities in a single step.

Methods

CMAPpro[3] consists of a stack of Neural Networks NN^k . Each network in the stack is a standard three-layer feed-forward network trainable by back propagation, and all the networks share the same topology: same input size, same hidden layer size, with one single output, which represents the residue-residue contact probability. Each level NN^k in the stack produces a contact map prediction, which is fed in input, and refined, in the successive level. Thus, the stack architecture allows to take into account the spatial correlation between residue-residue contacts, and it is used to organize the prediction is such a way that each level of the stack is meant to refine the prediction produced by the previous level. Other than the predictions coming from the previous level, the NN^k input includes three types of *fixed* information across the different levels:

Residue-residue features. These are the three most common type of features used for contact prediction. (1) Evolutionary information in form of sequence profiles, obtained with PSI-BLAST[1]. (2) Predicted secondary structure, obtained with SSpro[7]. (3) Predicted solvent accessibility, obtained with ACCpro[8].

Coarse features. We use a recurrent neural network (RNN) to predict coarse contact probabilities and orientation between secondary structure elements. In particular, the RNN is used to predict whether two secondary structure elements are in parallel contact, antiparallel contact, or no-contact. Such contact and orientation probabilities are fed into the network input.

Alignment features. We use an energy-based method to assign energies and then probabilities to the possible alignments between contacting secondary structure elements. The alignment probabilities provide some estimation of the possible spatial arrangement of two secondary structure elements. From such alignment probabilities we derive approximate probabilities of contact at the residue level, which are fed into the network input.

CMAPpro is trained on a large set of 2,356 non-redundant protein domains with less than 20% pairwise sequence identity, obtained from ASTRAL[2] release 1.73. The dataset of protein examples has been partitioned into 10 disjoint groups, so that no domains from two distinct groups belong to the same SCOP fold. Model training is performed using a standard 10-fold cross-validation procedure. The final CMAPpro is thus an ensemble of the ten distinct predictors.

Results

The performances of the deep architecture alone (with residue-residue features but without coarse and alignment features) have been tested in comparison to those of Neural Network (NN) and Recurrent Neural Network (RNN), trained from scratch on exactly the same data[3]. Experimental results show that the deep architecture leads to an accuracy of about 30% for long range contacts, roughly 10% above those of NN and RNN methods. The performances of CMAPpro[4] have been assessed on a large set of *new-fold* domains with respect to the training structures, as well as on the set of protein domains used for contact prediction in the two most recent CASP8[5] and CASP9[6] experiments. On these datasets, the accuracy of CMAPpro for long range contacts is close to 30%.

Availability

CMAPpro is available as part of the SCRATCH suite at: http://scratch.proteomics.ics.uci.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. Chandonia,J.M., Hon,G., Walker,N.S., Lo Conte,L., Koehl,P., Levitt, M., Brenner, S.E. (2004) The ASTRAL Compendium in 2004. *Nucl. Acids Res.*, **32**(1), D189-D192.
3. Di Lena, P., Nagata, K, Baldi, P. (2012) Deep Architectures for Protein Contact Map Prediction, *Bioinformatics*. In press.
4. Di Lena, P., Nagata,K., Baldi,P. Deep Spatio-Temporal Architectures and Learning for Protein Structure Prediction. *NIPS 2012 : Neural Information Processing Systems Conference*. Accepted for presentation.
5. Ezkurdia,I., Graña,O., Izarzugaza,J.M., Tress,M.L. (2009) Assessment of domain boundary predictions and the prediction of intramolecular contacts in CASP8. *Proteins*, **77**(9), 196-209.
6. Kryshtafovych,A., Fidelis,K., Moult,J. (2011) CASP9 results compared to those of previous CASP experiments. *Proteins*, **79**(10), 196-207.
7. Pollastri,G., Przybylski,D., Rost,B., Baldi,P. (2002) Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles, *Proteins*, **47**(2), 228-235.
8. Pollastri,G., Baldi,P., Fariselli,P., Casadio,R. (2002) Prediction of Coordination Number and Relative Solvent Accessibility in Proteins. *Proteins*, **47**(2), 142-153.

Tertiary Structure Predictions using the IntFOLD Server

L.J. McGuffin[1], D.B. Roche[2,3,4] and M.T. Buenavista[1,5,6]

[1] - *School of Biological Sciences, University of Reading, Reading, UK,*

[2] - *Commissariat à l'énergie atomique et aux énergies alternatives, Institut de Génomique, Genoscope, 2 rue Gaston Crémieux, 91057, Evry, France*

[3] - *Centre National de la Recherche Scientifique, UMR 8030, 2 rue Gaston Crémieux, 91057, Evry, France,*

[4] - *Université d'Evry-Val-d'Essonne, Boulevard François Mitterrand, 91025, Evry, France,*

[5] - *Biocomputing Section, Medical Research Council Harwell, Harwell Oxford, Oxfordshire, UK and*

[6] - *Beamline B23, Diamond Light Source, Didcot, UK.*

l.j.mcguffin@reading.ac.uk

Automated Tertiary Structure (TS) predictions were submitted using the TS component of the IntFOLD server [1] (IntFOLD-TS). Despite being a single template modelling method, the server performed well during the last round of CASP (CASP9). In particular, the B-factor scores that were assigned for each residue in a model provided an accurate reflection of observed local model quality[2]. The IntFOLD TS predictions were again included in CASP10 to provide a benchmark for monitoring performance of newer methods, such as IntFOLD2.

Methods

The IntFOLD method was essentially unmodified since CASP9; only the template and sequence databases were updated. The method works by integrating the alignment output from the SP3[3], SPARKS[3], HHsearch[4] and COMA[5] methods and then generating around 40 alternative single template based 3D models using Modeller[6]. For each target, all the generated models were then ranked using the ModFOLDclust2 QA method[7] and the top 5 were submitted.

The method included per-residue accuracy predictions in coordinate files, which were found to be accurate during the CASP9 experiment according to official assessments[8]. However, for CASP9, we did not make any attempt to correct the errors identified in single-template models. The method has since been updated for CASP10 and it now includes multi-template modeling guided by local quality assessment (see our IntFOLD2 abstract for more details).

Availability

The IntFOLD server is available at:
http://www.reading.ac.uk/bioinf/IntFOLD/IntFOLD_form.html.

1. Roche, D. B., Buenavista, M. T., Tetchner, S. J. & McGuffin**,** L. J. (2011) The IntFOLD server: an integrated web resource for protein fold recognition, 3D model quality assessment, intrinsic disorder prediction, domain prediction and ligand binding site prediction. *Nucleic Acids Res*. **39**, W171-6.
2. McGuffin, L. J. & Roche, D. B. (2011) Automated tertiary structure prediction with accurate

local model quality assessment using the IntFOLD-TS method. Proteins: Structure, Function, and Bioinformatics, **79 Suppl 10**, 137-46.
3. Zhou,H. & Zhou,Y. (2005) SPARKS 2 and SP3 servers in CASP6. *Proteins*. **61** (S7), 152-156.
4. Söding, J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*. **21**, 951-96.
5. Margelevičius,M. & Venclovas Č. (2010) Detection of distant evolutionary relationships between protein families using theory of sequence profile-profile comparisons. *BMC Bioinformatics*. **11**, 89.
6. Sali,A. & Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol*. **234**, 779-815.
7. 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.
8. Mariani V., Kiefer F., Schmidt T., Haas J. & Schwede T. (2011) Assessment of template based protein structure predictions in CASP9. *Proteins*. **79 Suppl 10**, 37-58.

Fully Automated Prediction of Tertiary Structures, Intrinsic Disorder and Binding Site Residues Using the IntFOLD2 Server

L.J. McGuffin[1], D.B. Roche[2,3,4] and M.T. Buenavista[1,5,6]

[1] - *School of Biological Sciences, University of Reading, Reading, UK,* [2] - *Commissariat à l'énergie atomique et aux énergies alternatives, Institut de Génomique, Genoscope, 2 rue Gaston Crémieux, 91057, Evry, France,* [3] - *Centre National de la Recherche Scientifique, UMR 8030, 2 rue Gaston Crémieux, 91057, Evry, France,* [4] - *Université d'Evry-Val-d'Essonne, Boulevard François Mitterrand, 91025, Evry, France,* [5] - *Biocomputing Section, Medical Research Council Harwell, Harwell Oxford, Oxfordshire, UK and* [6] - *Beamline B23, Diamond Light Source, Didcot, UK.*

l.j.mcguffin@reading.ac.uk

The IntFOLD2 server integrates our latest methods for: fold recognition, domain boundary prediction, prediction of intrinsically disordered regions, prediction of protein-ligand interactions and the global and local quality assessment of predicted 3D models of proteins. Our main focus for IntFOLD2 was the improvement of 3D protein models using multiple templates guided by single-template model quality assessment.

Methods

For CASP10, a bespoke version of the server was developed in order to simultaneously return results for 3 prediction categories: tertiary structure (TS), disorder (DR), and ligand binding site residues (FN). Local quality assessment results were returned as scores in the B-factor column of each TS model file. Full QA results were returned by separate servers (see our ModFOLD4 and ModFOLDclust abstracts for details).

TS predictions: Our new TS method was developed with the aim of fixing local errors, identified in an initial pool of single template models, through iterative multi-template modeling. The method attempts to exploit our CASP9 success in local quality prediction[1] by taking the per-residue errors into consideration during multiple alignment selection.

In a recent paper, we compared several alternative alignment section methods for multi-template modeling[2]. We discovered that using accurate local model quality scores to guide the alignment selection was the most consistent way to significantly improve models for each of the sequence to structure alignment methods tested. In addition, using accurate global model quality for re-ranking alignments, prior to selection, further improved the majority of the multi-template modeling approaches that we tested. Furthermore, subsequent clustering of the resulting population of multiple-template models significantly improved the quality of selected models compared with the previous version of our tertiary structure prediction method (IntFOLD-TS).

For the IntFOLD2 server TS predictions, nine different fold recognition methods were installed and run in-house to generate up to 10 sequence-to-structure alignments each, resulting in up to 90 alternative single-template based models being generated for each CASP target. The fold recognition methods that we used were SP3[3] and SPARKS2[3], HHsearch[4], COMA[5] and the

5 alternative threading methods that are integrated into the current LOMETS package[6] - QQQ, GGGd, GGGf, NNNd and SSSc.

In the first iteration of the IntFOLD2 TS method, all single-template models were assessed using ModFOLDclust2 to assign global and local model quality scores. Using the single template model quality scores and other criteria, alignments were selected from each fold recognition method and used to build multiple-template models, with the aim of minimizing errors. The multi-template modeling alignment selection methods resulted in the generation of a new population of models for each target. In the second iteration the new multi-template models were then assessed using ModFOLDclust2[7] and the top-ranked models were designated as the IntFOLD2 TS predictions.

DR predictions: The latest version of our DISOclust method[8] was used to generate automated DR submissions via the IntFOLD2 server. The new method uses the ModFOLDclust2 QMODE2 output in order to identify the regions of high variability occurring in the 3D models generated by the IntFOLD2 TS method.

FN predictions: The latest version of FunFOLD[9] was used to generate automated FN submissions via the IntFOLD2 server. The method uses structural superpositions of the top ranked IntFOLD2 3D models and related templates with bound ligands in order to identify putative contacting residues. The method uses a hierarchical agglomerative clustering approach for ligand identification and residue selection.

Availability

The IntFOLD2 server is available at:
http://www.reading.ac.uk/bioinf/IntFOLD/IntFOLD2_form.html.

1. McGuffin, L. J. & Roche, D. B. (2011) Automated tertiary structure prediction with accurate local model quality assessment using the IntFOLD-TS method. Proteins: Structure, Function, and *Bioinformatics*. **79 Suppl 10**, 137-46.
2. Buenavista, M. T., Roche, D. B. & McGuffin, L. J. (2012) Improvement of 3D protein models using multiple templates guided by single-template model quality assessment. *Bioinformatics*. **28**, 1851-1857.
3. Zhou,H. & Zhou,Y. (2005) SPARKS 2 and SP3 servers in CASP6. *Proteins*. **61** (S7), 152-156.
4. Söding, J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*. **21**, 951-96.
5. Margelevičius,M. & Venclovas Č. (2010) Detection of distant evolutionary relationships between protein families using theory of sequence profile-profile comparisons. *BMC Bioinformatics*. **11**, 89.
6. Wu, S. and Zhang, Y. (2007) LOMETS: A local meta-threading-server for protein structure prediction. *Nucleic Acids Research*. **35,** 3375-3382.
7. 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.
8. McGuffin, L.J. (2008) Intrinsic disorder prediction from the analysis of multiple protein fold recognition models. *Bioinformatics*. **24**, 1798-1804.

9. Roche, D. B., Tetchner, S. J. & McGuffin, L. J. (2011) FunFOLD: an improved automated method for the prediction of ligand binding residues using 3D models of proteins. *BMC Bioinformatics*. **12**, 160.

Jiang_Fold: A Packing Cluster-based Fold Recognition Server

Aiping Wu*, Lizong Deng*, Tingrui Song*, Wentao Dai, Zhichao Miao, Xuan Wang, Taijiao Jiang#

Key Laboratory of Protein and Peptide Pharmaceuticals, Institute of Biophysics, Chinese Academy of Sciences, Beijing, China, 100101

*Authors contributed equally to this work

#taijiao@moon.ibp.ac.cn

To predict structures for all server targets in CASP10, we developed an automatic fold recognition method named Jiang_Fold. Different from our previous fold recognition method FR-t5 [1], Jiang_Fold is based on a new structural descriptor, packing cluster (PC) which is represented as a buried highly connected side-chains network enclosed by main-chains. PC, which can be regarded as a type of 'structural code', was used to guide sequence-structure alignment in fold recognition, leading to an improvement on the performance of fold recognition especially for remote homology. We further improved the threading speed and accuracy by implementing an iterative threading strategy.

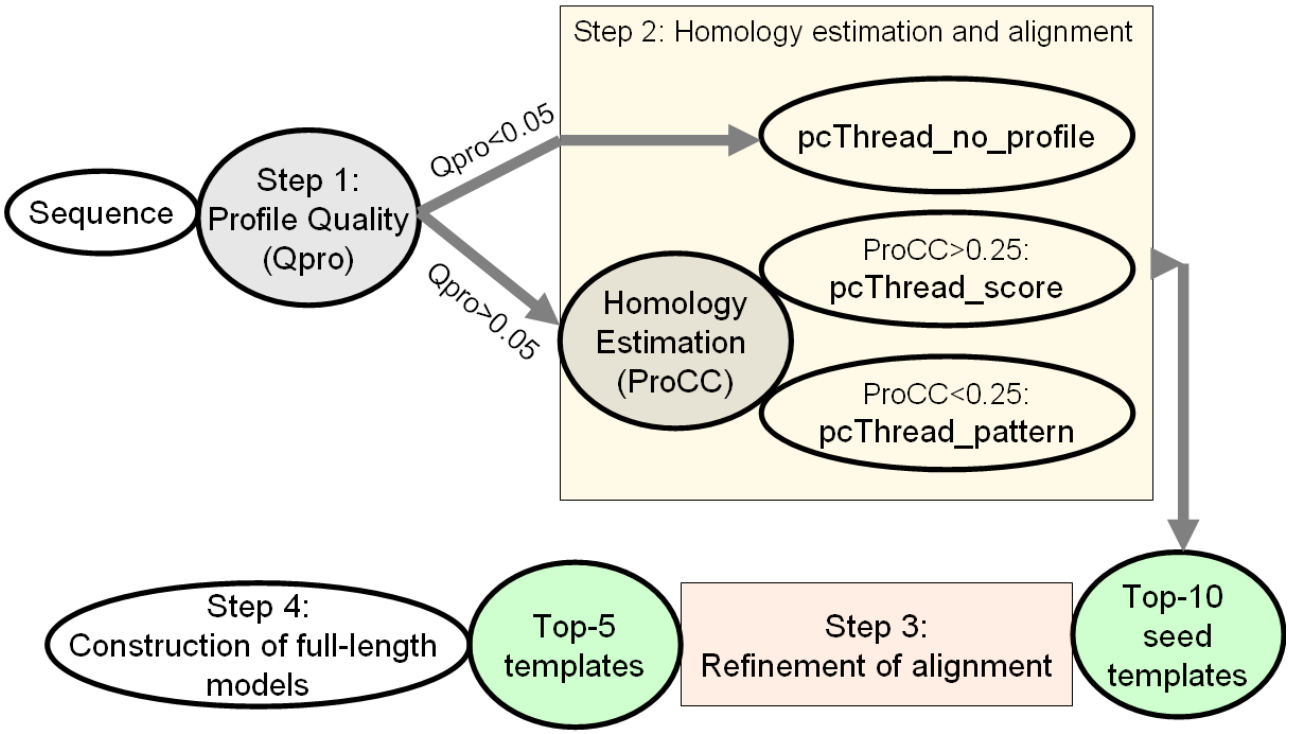


Figure 1: The framework of fold recognition server Jiang_Fold for CASP10.

The framework of Jiang_Fold consists of four key steps (Figure 1) described as follows:

Step 1: Quality Evaluation of Sequence Profile

For each target sequence to be modeled, we first evaluate its sequence profile quality using the method of Qpro (Quality of Profile) we recently developed.

Step 2: Sequence-template Alignment and Homology Estimation
If Qpro<0.05, pcThread_no_profile module was used for sequence-template alignment which does not consider profile term in alignment scoring function. Otherwise, pcThread_score module was used. The sequence-template is performed against a template library of SCOP <= 90% sequence similarity [2]. Meanwhile, we calculated the correlation coefficient between profiles of target sequence and fold template (ProCC). If ProCC<0.25, which means low homology between the target and template, we re-did sequence-template alignment using pcThread_pattern module that was trained on low homology targets. All the sequence-template alignments ware evaluated by Z-scores and the top 10 alignments with highest Z-scores were chosen as seeds for next round of threading procedure.

Step 3: Refinement of sequence-template alignment
To improve the performance of fold recognition, we obtained all structures with same folds from the complete SCOP template library for the top 10 seed templates that were generated above. Then target sequence was aligned with all these structures including the top 10 seed templates. Finally the top 5 structural templates with highest Z-scores were obtained.

Step 4: Construction of Full-length Model
For target sequences of multiple domains, the templates generated could only cover partial protein since the domain-based template library is used. Therefore, we need to check whether the sequence-template alignment covers the whole target sequence. If not, we generated templates to cover different domains of the target sequence, and used PatchDock [3] to construct the full-length model.

1. Hu Y, Dong XX, Wu AP, Cao, Y, Tian LQ, Jiang TJ. (2011). Incorporation of local structural preference potential improves fold recognition, PLoS One. 6(2): e17215.
2. 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.
3. Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ. (2005). PatchDock and SymmDock: servers for rigid and symmetric docking. Nucl. Acids. Res. 33: W363-367.

Jiang_Server: an integrated protein structure modeller

Wentao Dai*, Zhichao Miao*, Lizong Deng, Tingrui Song, Xuan Wang, Aiping Wu, Taijiao Jiang#

Key Laboratory of Protein and Peptide Pharmaceuticals, Institute of Biophysics, Chinese Academy of Sciences, Beijing, 100101, China;
*Authors contributed equally to this work
#taijiao@moon.ibp.ac.cn

The protein structure prediction server, Jiang_Server was developed by integrating our threading program FR-t5[1] and de novo fragment-based assembly program (denoted as NCACO-assembler) using NCACO-score function[2]. The predicted models were ranked by a model selection method that we recently developed[3]. The framework of Jiang_Server is depicted as Figure 1.

Methods:

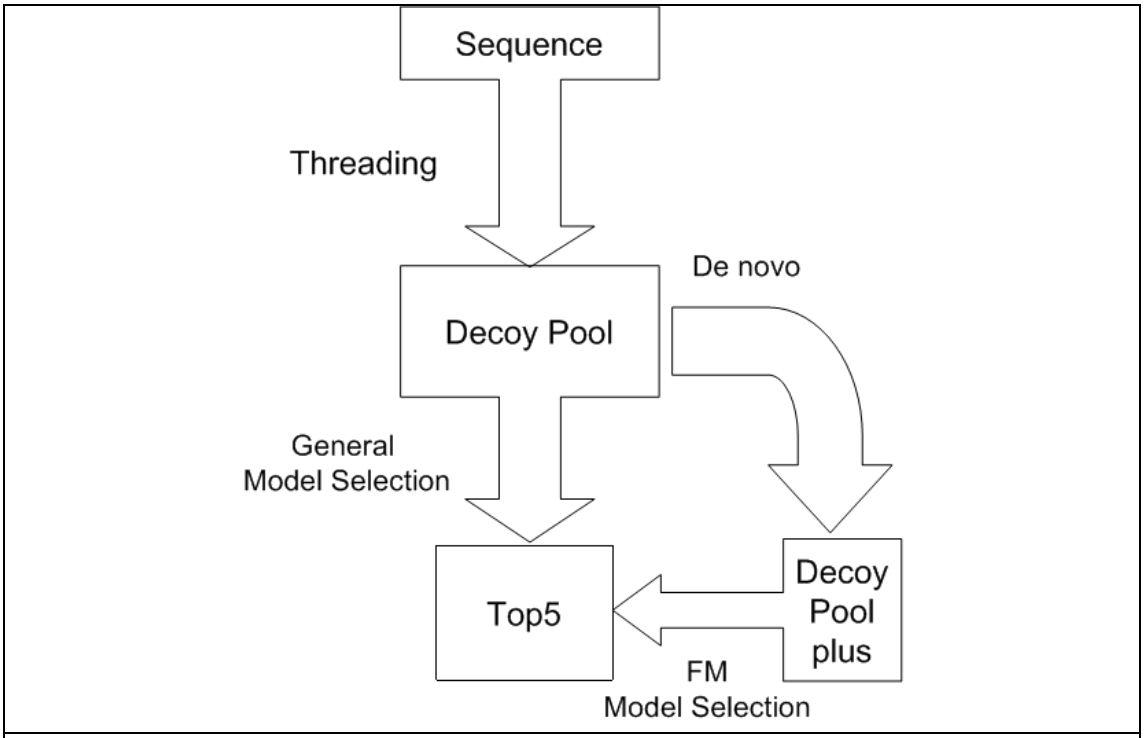


Figure 1. Framework of the integrated protein structure modeller Jiang_Server.

Jiang_Sever was used to generate structure models for the target sequences in the 10th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP10). For each target sequence, we first used the threading program FR-t5 to generate 50 structure models. Then, a model selection program was employed to assess these structure models, whose quality was measured by the Gscores (General Model Selection). If there existed ≥5 models with Gscores ≥2.95, the top 5 models were selected as the final prediction models. Otherwise, ten more models were generated by the de novo assembly program NCACO-assembler, and the top 5 models with highest Pscores (FM Model Selection) were selected from these 60 structure models. General Model Selection and FM Model Selection

are all based on SVR(support vector machine regression) algorithm, while Pscore has better performance than Gscore in hard targets for its training process.

Results

The performance of Jiang_Server was evaluated on the 117 targets from CASP9. To mimic the participation in CASP9, the PDB database used to generate fragment libraries and the nr database used to generate sequence profiles were constructed based on the data before the start of CASP9 (namely Apr 14, 2010). In order to assess the performance of Jiang_Server, we compared it to our previous prediction server Jiang_THREADER. As shown in Table 1, Jiang_Server shows a better performance than Jiang_THREADER.

	ALL(117 targets)	Hard Targets(29)
Jiang_Server	75.42	7.92
Jiang_THREADER	73.78	6.20

Table 1. Comparison of Jiang_Server and Jiang_THREADER in CASP9 targets. The scores are the total TM-scores over Top1 models for all targets or hard targets.

1. Tian,L. Wu,A. Cao,Y. Dong,X. Hu,Y. and Jiang,T. (2011). NCACO-score: an effective main-chain dependent scoring function for structure modeling. BMC bioinformatics 12, 208-230.
2. Hu,Y. Dong,X. Wu,A. Cao,Y. Tian,L. and Jiang,T. (2011). Incorporation of Local Structural Preference Potential Improves Fold Recognition. PLoS ONE 6, e17215.
3. Work in preparation for publication.

Protein structure prediction using pGenTHREADER and FRAGFOLD

D.W.A. Buchan[1], D. Cozzetto[1], Tim Nugent[1], Tomasz Kosciolek[1], Stuart Tetchner[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 prediction efforts were aimed at generating models for the harder prediction targets; if there were obvious homologous matches with target domains a simple meta prediction method based on our hGen3D method was used (see hGen3D/NewSerf server abstract elsewhere).

Methods

For CASP10, for those target domains which we believed could not be reliably predicted using fold recognition methods (such as pGenTHREADER[1]), we used FRAGFOLD [2] to generate up to 5 structures. This approach to protein tertiary structure prediction is based on the assembly of recognized supersecondary structural fragments taken from highly resolved protein structures using a simulated annealing algorithm. The current release of FRAGFOLD, version 4.7, was very similar to the version used in CASP9. As many as 5,000 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.

For 12 targets, we attempted to predict residue-residue contacts using PSICOV [5], which identifies correlated mutations by applying a graphical lasso procedure to large multiple sequence alignments. Unfortunately none of these targets turned out to be de novo targets, and so we are unable to evaluate the efficacy of this approach in the CASP10 experiment. The lack of homologous sequences for CASP10 targets seems rather unusual based on general family size statistics. Whether this indicates a bias in the selection of CASP targets, or perhaps a bias in target selection in the major structural genomics consortia, remains to be seen.

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. Prior to submission DaliLite was used with the model in case it could be used to find a distant homologous structure which could be used to alternatively model the target sequence.

Results

Predictions of folds were submitted for all targets. Due to a lack of homologous sequences for the de novo targets, it is not possible to evaluate the prospects of using newly developed contact prediction methods for advancing the state-of-the-art in de novo protein structure prediction.

1. 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.
2. 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.

3. Pettitt CS, McGuffin LJ, Jones DT. (2005) Improving sequence-based fold recognition by using 3D model quality assessment. Bioinformatics. 21, 3509-3515.
4. Zhang Y, Skolnick J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 22, 2302-2309.
5. Jones, D.T., Buchan, D.W., Cozzetto, D. & Pontil, M. (2012) PSICOV: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics. 28, 184-190.

Model selection and refinement

C. Keasar

Ben Gurion University of the Negev
Chen.keasar@gmail.com

In the current CASP experiment we focused on model selection and refinement, using server models as our starting point for all simulations.

For model refinement we used energy optimization with a cooperative energy function, as we did in previous CASP experiments. This time, however, we used a different scoring function for model ranking and selection. This scoring function is applicable to energy-optimized models considering the energy terms, but weighs them differently and complements them with evolutionary- based information in the Spirit of Kalman *et al*[1]

Methods

We took part in three CASP10 categories: QA, refinement, and tertiary structure prediction. For the three tasks we used the same protocol, presented below. QA required steps 1-3 of the protocol, for refinement we used steps 5-8, and the whole protocol was used for structure prediction.

1. Server models were downloaded from the CASP web site.
2. All models were energy minimized to reduce server-specific characteristics that mask quality related differences.
   a. Our energy function requires a secondary structure assignment. To this end we used two alternative secondary structure predictions: PSI-PRED[2] and SAM-T08[3]. Thus each model was minimized twice, each time with a different secondary structure assignment.
   b. It should be noted that quite a few of the server models did not pass this stage. Typically, this had to do with some (often minor) structural distortion in the model that rendered the minimization numerically unstable. These models were not considered further.
3. Models were ranked by a scoring function. This function was trained with native secondary structures, and turned out to be very sensitive to the secondary structure assignment and often provided different rankings for the two choices of secondary structure assignments. In our work for CASP10 we chose for each target the secondary structure (and thus ranking) that received the higher score.

   At this stage QA predictions were submitted. Structure modeling targets continued to the next stage.
4. We visually inspected the top ranking models (around 20 models depending on time limitations). Visual inspections had two roles:
   a. Removal of identical models and obvious outliers (i.e. apparent FM models in clear TBM targets).
   b. Identification of multi-domain targets. Typically the top ranking models agreed on domain boundaries. In cases of doubt we consulted BLAST[4] and 3D-Jury[5]. In what follows, identified domains were handled separately and submitted as independent fragments.

   Here refinement targets entered the protocol.

5. The selected models were often manually manipulated. Most of these manipulations were limited to changes in the secondary structure assignment and tethering of conserved structural elements. In a few cases, when distortions seemed obvious and time permitted, actual structural changes were made.
6. Models were energy optimized using MCM[6]
7. Optimized models were ranked by the scoring function and submitted.

Two central elements of the above scheme are the energy and scoring functions. The energy function includes non-cooperative torsion angle[7] and atom-pair potentials[8], and cooperative meta-terms that bound the latter from reaching values that are too low[9]. In addition the energy function includes a cooperative hydrogen bonding term[10], and a cooperative solvation term[9]. The scoring function includes some of the above energy terms as well terms that penalize exposed conserved residues, and a term that penalizes deviations from the assigned secondary structure. The weights of these terms were learnt from a subset of the CASP8 server models.

Results

According to our self-evaluation, given the available native structure as of Sept. 28th :

1. In the QA category we have reached weighted average correlations[11] of 5.2 and 2.6 for the stage1 and stage2 tracks, respectively.
2. In the refinement category we slightly improved (model #1) five out of ten models in terms of GDT_TS.
3. No evaluation of tertiary structures was performed yet.

Availability

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

1. Kalman, M. & Ben-Tal, N. (2010). Quality assessment of protein model-structures using evolutionary conservation. *Bioinformatics* **26**, 1299–1307
2. Jones, D.T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology* **292**, 195–202
3. Katzman, S., Barrett, C., Thiltgen, G., Karchin, R. & Karplus, K. (2008). Predict-2nd: a tool for generalized protein local structure prediction. *Bioinformatics* **24**, 2453–2459
4. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402
5. Ginalski, K., Elofsson, A., Fischer, D. & Rychlewski, L. (2003). 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics* **19**, 1015–1018
6. Li, Z. & Scheraga, H.A. (1987). Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *PNAS* **84**, 6611–6615
7. Amir, E.-A.D., Kalisman, N. & Keasar, C. (2008). Differentiable, multi-dimensional, knowledge-based energy terms for torsion angle probabilities and propensities. *Proteins: Structure, Function, and Bioinformatics* **72**, 62–73
8. Summa, C.M. & Levitt, M. (2007). Near-native structure refinement using in vacuo energy minimization. *PNAS* **104**, 3177–3182
9. Maximova, T., Kalisman, N. & Keasar, C. Unpublished.
10. Levy-Moonshine, A., Amir, E.D. & Keasar, C. (2009). Enhancement of beta-sheet assembly by cooperative hydrogen bonds potential. *Bioinformatics* **25**, 2639–2645
11. Kryshtafovych, A., Fidelis, K. & Tramontano, A. (2011). Evaluation of model quality predictions in CASP9. *Proteins: Structure, Function, and Bioinformatics* **79**, 91–106

Prediction of protein structure with the use of UNRES force field with Conformational Space Annealing and Dynamic Fragment Assembly

Bartłomiej Zaborowski[1,2], Cezary Czaplewski[1], Adam Liwo[1], Juyoung Lee[2], Jooyoung Lee[2*]

[1] – *Faculty of Chemistry, University of Gdańsk, Sobieskiego 18, 80-952 Gdańsk, Poland*

[2] – *Korea Institute forAdvanced Studies, 207-43 Cheongnyangni 2-dong, Dongdaemun-gu, Seoul 130-722, Korea*
jlee@kias.re.kr

In the last few years, we have been developing the physics-based UNRES[1] force field for coarse-grained simulations of protein structure and dynamics, which performs well in the physics-based prediction of protein structure. Because of still imperfect representation of local interactions, errors in secondary structure, local details, and packing are sometimes substantial. Recently, the Dynamic Fragment Assembly (DFA) method was proposed by Sasaki et el.[2,3], in which knowledge-based information of a given protein is incorporated into a set of local potentials. The DFA method was subsequently implemented by Lee et al.[4] in protein-structure prediction by global conformational search with Conformational Space Annealing (CSA)[5] and the CHARMM2 force field. The purpose of this exercise was to assess how does this supplementary information of local interactions improve the performance of UNRES.

Methods

In the UNRES model[1], 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 cluster-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.

The DFA routines by Sasaki et al.[2,3] were incorporated into the UNRES package. For each target, the knowledge-based potentials, which consisted of virtual-bond-angle and virtual-bond-dihedral-angle biasing potentials, biasing potentials imposed on local distances (within 9-residue fragments), neighbor-number potentials, and a biasing potential accounting for β-sheet propensity, were constructed with the use of the procedure developed by Sasaki et al.[2,3] Subsequently, the Conformational Space Annealing (CSA)[5] global-optimization runs were carried out with UNRES+DFA as the target function and cluster analysis of the results was carried out by the Ward minimum-variance method[6] to identify conformational patterns. The clusters were ranked according to the UNRES+DFA energy of the lowest-energy member and the lowest energy conformation of a cluster was selected to represent the entire cluster.

The representatives of five lowest-energy clusters were subsequently selected for further stages of the procedure. Then, all-atom structures were constructed from the UNRES structures by using our physics-based procedure[7,8] for the reconstruction of all-atom backbone and all-atom side chains. In this procedure, the peptide groups are positioned first by Monte Carlo optimization of the sum of dipole-dipole interaction energy and energy of backbone-local interactions, subject to the Cα-trace geometry resulting from coarse-grained simulations and, subsequently, side chains are added subject to the condition of minimum overlap[8]. Finally, the

structures are refined by energy minimization with the ECEPP/3 force field[9].

Availability

The UNRES package to perform coarse-grained simulations is available at http://www.unres.pl

1. Liwo. A., Czaplewski, C., Ołdziej, S., Rojas, A. V., Kaźmierkiewicz, R., Makowski, M., Murarka, R. K. & Scheraga, H. A. (2008). Simulation of protein structure and dynamics with the coarse-grained UNRES force field. In: *Coarse-Graining of Condensed Phase and Biomolecular Systems.*, ed. G. Voth, Taylor & Francis, 2008, Chapter 8, pp. 107-122.
2. Sasaki, T. N. & Sasai, M. (2004). A coarse-grained langevin molecular dynamics approach to protein structure reproduction. *Chemical Physics Letters* **402**, 102-106.
3. Sasaki, T. N., Cetin, H. & Sasai, M. (2008). *Biochemical and Biophysical Research Communications* **369**, 500–506.
4. Lee, J., Lee, J., Sasaki, T. N., Sasai, M., Seok, C. & Lee, J. (2011). De novo protein structure prediction by dynamic fragment assembly and conformational space annealing. *Proteins: Structure Function and Bioinformatics* **79**, 2403-2417.
5. Lee, J., Scheraga, H. A. & Rackovsky, S. (1997). *Journal of Computational Chemistry* **18**, 1222-1232.
6. Murtagh, F. & Heck, A. (1987) *Multivariate Data Analysis*, Kluwer Academic Publishers.
7. Kaźmierkiewicz, R., Liwo, A. & Scheraga, H. A. (2002). Energy-based reconstruction of a protein backbone from its α-carbon trace by a Monte-Carlo method. *Journal of Computational Chemistry* **23**, 715-723.
8. Kaźmierkiewicz, R., Liwo, A. & Scheraga, H. A. (2003). Addition of side chains to a known backbone with defined side-chain centroids. *Biophysical Chemistry* **100**, 261-280.
9. Nemethy, G., Gibson, K. D., Palmer, K. A., Yoon, C. N., Paterlini, G., Zagari, A., Rumsey, S. & Scheraga, H. A. (1992). Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm with application to proline-containing peptides. *The Journal of Physical Chemistry* **96**, 6472-6484.

Structure Prediction and Refinement Using Constraints-Based CABS Model and Knowledge-Based Secondary Structural Fragments Interaction and Residue Environmental Potential

Hyungrae Kim[1], Andrzej Kolinski[2], & Daisuke Kihara[1]

[1] – *Department of Biological Science, Computer Science, Purdue University, West Lafayette, IN, USA*

[2] – *Department of Chemistry, Warsaw University, ul. Pasteura 1, Warszawa, Poland*

dkihara@purdue.edu

We submitted structure models in three tertiary structure prediction categories in CASP10. They are regular (TS), refinement (TR), assisted (Tc) structure prediction targets of CASP10. 255 models for 51 regular human-server targets, 133 models for 27 refinement targets, and 75 models for 15 assisted prediction targets were submitted in total.

Methods

Our structure prediction scheme is based on the CABS lattice model-based ab initio folding method [1]. We made a number of significant modifications to the original CABS to allow more flexible move during Monte Carlo (MC) structure optimization and for better scoring of conformations. Moreover, we enhanced the Replica Exchange Monte Carlo (REMC) scheme to incorporate predicted local quality information of replica structures. We have also newly implemented novel secondary structural fragment-fragment interaction potential into the CABS force field. The model selection was performed by applying a novel residue-level environment potential. Our interest and the novel implementations are mainly for improving packing patterns of secondary structural fragments and side-chains of neighboring residues.

Below is the summary of our prediction procedure using CABS:

1. Modification of CABS for reflecting expected local errors.

To allow more flexible move in CABS, chain moves on the 3D lattice are modified. The lattice space as well as MC move parameters were modified. In addition, a new REMC routine was employed to be able to consider predicted local errors during folding process. In principle, MC moves are modulated to enhance moves at local regions that are predicted to be of poor quality.

2. Novel secondary structural fragment-fragment interaction potential

On top of the CABS force field, we implemented a novel secondary structural fragment-fragment interaction potential. The knowledge-based potential is dependent on fragment crossing angle and the distance of fragments. The potential was incorporated into CABS force field and this fragments interaction potential was weighted as being half of the total potential value of a protein conformation. A several different potentials were prepared based on the mutual distance between fragments. The fragment-fragment interaction potential was shown to be very effective in making large fragment-based motion and achieving correct fragment packing observed in native structures.

3. Starting Replica structures to CABS

We generated and selected about 30 starting replica models. Some of them are taken from server predictions. We also used our in-house fold quality assessment, Sub-AQUA [2]. After removing low quality models that are very different from the others, on average 20-26 different models were finally used as initial replica structures in a CABS REMC run.

4. Running CABS

In total 12 independent CABS were run simultaneously. We iterated the running of CABS at least 2 to 3 times and continued the iteration until the 12 structures from the largest cluster of every 12 CABS runs nearly converged to a few similar structures. Restraints were applied during the simulation to regions that are consistent among server models.

5. Model selection by a novel residue environmental potential

A novel residue environmental potential was developed for model selection. The potential was used to select best models from the whole Monte-Carlo trajectories. This potential essentially examines if side-chain environment at each residue in a model can find similar environment in known structures in a structure database or not. The environment of a residue is defined by the types and the number of neighboring residues. A residue in a model is scored by counting the number of similar environments observed in the representative set of protein structures. Our residue environment potential was shown to be superior to several existing scoring functions in selecting near native decoys in our benchmark test (manuscript in preparation). Selection of models was performed by applying the residue environment potential to models followed by manual inspection. Best five models based on the environment potential score were selected from the whole MC trajectories.

Summary

In this work, original CABS ab initio folding program was significantly enhanced in three aspects: First, more flexible moves were made possible to explore a larger and more protein-like conformational space. Second, the assessed quality of local structures are explicitly considered and implemented in the MC optimization. Third, two novel potentials, one for fragment-fragment interactions and another one for side-chain environment were developed and implemented. These two potentials are complementary to each other but aimed for the same purpose of improving packing patterns of fragments and residues.

Acknowledgements

This work has been supported by grants from the National Institutes of Health (R01GM075004, R01GM097528), National Science Foundation (EF0850009, IIS0915801, DMS0800568), and National Research Foundation of Korea Grant funded by the Korean Government (NRF-2011-220-C00004).

1. Kolinski, A. (2004). Protein modeling and structure prediction with a reduced representation. Acta Biochimica Polonica 51: 349-371
2. Yang YD, Spratt P, Chen H, Park C, Kihara D. (2010). Sub-AQUA: Real-value quality assessment of protein structure models. 23: 617-632.

Kloczkowski_Lab

Eshel Faraggi, Andrzej;Kloczkowski

Battelle Center for Mathematical Medicine, NCH

Protein structure prediction can be separated between two tasks: sample the configuration space of the protein chain, and assign a fitness between these hypothetical models and the native structure of the protein. One of the more promising developments in this area is that of knowledge based energy functions. However, standard approaches using pair-wise interactions have shown shortcomings demonstrated by the superiority of multi-body-potentials. These shortcomings are due to residue pair-wise interaction being dependent on other residues along the chain. We developed a method that uses whole protein information filtered through machine learners to score protein models based on their likeness to native structures.

Materials and Methods

Single chain models were collected from the PDB and any redundant sequences were removed. Additional models were collected from previous CASP experiments. For all models we calculated parameters associated with the distance to the solvent and with distances between residues. These parameters, in addition to energy estimates obtained by using a four-body-potential, DFIRE, and RWPlus were used as training for machine learners to predict the fitness of the models. For the human prediction portion of the CASP10 experiment we took the top 150 ranked server models as supplied by the organizers and ranked them according to our trained server. The top five ranked models were then submitted as our human prediction. No human intervention was carried out in ranking top models. During the CASP experiment we have discovered an over-training mistake that we absent mindedly committed and our model were slightly modified midway through the experiment. Local testing indicated however that the change in ranking coming about from the changing of models was minimal.

Discussion

Testing on CASP 9 targets showed that our method is superior to the common DFIRE and its derivatives as well as to the current version of RWPlus, both of which are considered a standard in the field. Further testing showed that our prediction were on par with the best ranking methods from all groups participating in the CASP9 experiment. These results are currently being improved and summarized into a paper. The server will be part of the bioinformatics services page of the Kloczkowski Lab at the Battelle center.

Iterative knowledge-based minimization protocol

G. Chopra and M. 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. Such potentials implicitly include the effects of solvent and the crystal environment as they are derived from exprimentally determined structures deposited in the PDB. Our knowledge-based potential energy surface has an attractor basin to move the near-native conformation (refinement targets) consistently closer to the native structure but the surface is very rugged with many local minima close to the native conformation[1] (Figure 1). An iterative scheme to perturb the confirmations slightly may work to move the conformations closer to the native conformation by escaping these rugged minima. To investigate this, we submitted predictions for all the targets for CASP-ROLL and CASP10 by processing them through three different pipelines. We test our consistent knowledge-based refinement protocol[2] for structure prediction and used an iterative version of the pipeline for the refinement targets. All steps in these pipelines are automated: if they prove useful, we hope to run them in the server category at future CASPs.

Methods

For CASP-ROLL targets we selected the server predicted models and used the protocol implemented in our KoBaMIN web server (http://csb.stanford.edu/kobamin/). KoBaMIN is a protein structure refinement server that employs a simple, accurate, consistent and computationally efficient protocol based on a knowledge-based potential energy function[1; 2; 3]. The refinement protocol involves a two-step process. First, the server uses ENCAD[4] to refine the protein by a highly convergent energy minimization algorithm with an all-atom knowledge-based potential of mean force that implicitly includes the effect of solvent, KB01[1; 2]. ENCAD's implementation of the KB01 potential enables rapid refinement of structures (less than 5 minutes for a protein of chain length 300), often bringing them closer to the true native conformation. Second, a restrained energy minimization is performed using MESHI[5] to correct side-chain rotamer positions and other details of the stereochemistry. This protocol has been tested extensively on all human and server models predicted in CASP7 and performed consistently well[2]. The KoBaMIN server protocol was also used as the "end" step by many groups for CASP10, specifically, by the WeFold initiative.

For the template-based and template-free modeling category, we selected the top five server models using our model selection protocol as implemented in BITS server for stage-2 models as explained in the abstract of the BITS group. The BITS server is based on the simple premise that the functional sites in protein are more conserved than their global structure, and that the quality of local structure in predicted models can be used as a measure of overall structural quality. Then we used our KoBaMIN protocol[3] on the top five selected server models and submitted them as predictions.

For the refinement targets we followed an iterative KoBaMIN protocol. In order to escape from local minima and move closer to the native structure, the starting model was minimized through the KoBaMIN protocol and the resulting model was again processed by the

same protocol. This iteration was done five times. As the KoBaMIN protocol consists of KB01 energy minimization step followed by MESHI side-chain rotamer position correction with restrained backbone flexibility[2; 3], such an iterative scheme is effective. For refinement, we used the starting model provided and did not use information about problematic regions in this model. Our protocol was applied to the entire structure with no 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. The five iterated models were submitted as predictions.

Availability

The KoBaMIN protocol[3] is available as an online server at http://csb.stanford.edu/kobamin/. It calculates Cα RMSD, 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 web servers in the future and hope to run them under the server category at all future CASPs.

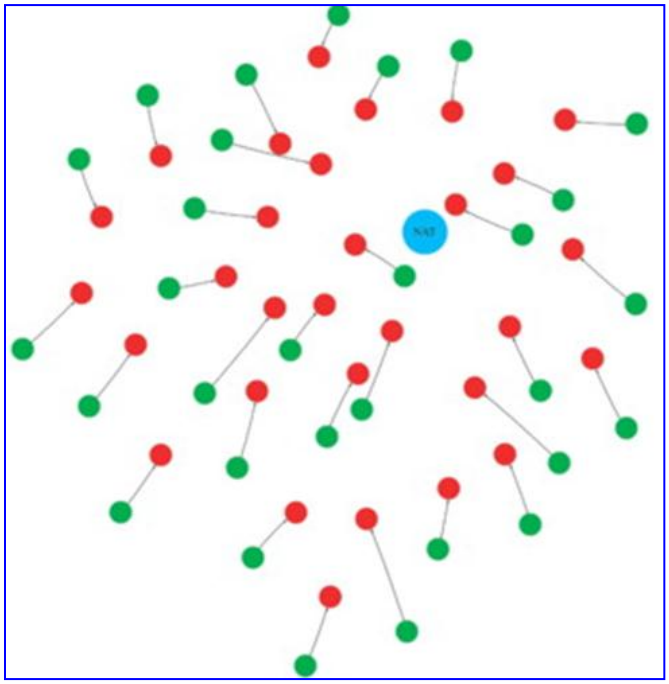


Figure 1 (modified from previous work[1]). Comparing the directed movement on the potential energy surface for energy minimization with KB potential. The starting protein structures are green points, the native structure is the cyan disk and the final energy minimized structures are shown as red points. This projection of the multidimensional energy surface is made using a 61 × 61 matrix of pairwise RMS values (30 initial, 30 final energy minimized, and the native structure). An attractor basin is seen here but the energy surface contains many local minima near the native state preventing it from being reached. Thus using a MESHI side-chain perturbation scheme with iterative KB minimization could help escape such minima and move the structures closer to the native conformation. The distance between any two points is scaled by how far apart the structures are in the Cα RMSD space.

1. Chopra, G., Summa, C. M. & Levitt, M. (2008). Solvent dramatically affects protein structure refinement. *Proc Natl Acad Sci USA*. **105**, 20239-20244.
2. Chopra,G., Kalisman,N. & Levitt,M. (2010). Consistent refinement of submitted models at CASP using a knowledge-based potential. *Proteins*. **78**, 2668-2678.
3. Rodrigues, J. P. G. L. M., Levitt, M. & Chopra, G. (2012). KoBaMIN: a knowledge-based minimization web server for protein structure refinement. *Nucleic Acids Res.* **40**, W323-W328.

4. Levitt M., Hirshberg M., Sharon R. & Daggett V. (1995) Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. *Comput Phys Commun*. **91**, 215-231.
5. 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.

Meld: modeling with limited data

J.L.MacCallum[1], A. Perez[1], A. Roy[1] and K.A. Dill[1]

[1] – *Laufer Center for physical and quantitative Biology, Stony Brook University*
justin.maccallum@me.com

In principle, physics-based methods can be used to fold proteins and predict their structure. However, it has remained an elusive task due to the vastness of the protein conformational space and the roughness of the potential energy surface. In the process of CASP10 we have been using a physics-based methodology that incorporates prior information to focus sampling on the relevant parts of conformational space. This prior information is encoded as restraints that are used as perturbations to the system's Hamiltonian using a procedure based on multiple replicas—the Hamiltonian Replica Exchange method[1]. We construct a ladder of replicas in which the temperature and the strength of the restraints change, and in which springs are selected from a pool of possible springs based on physical principles. At high temperatures, restraints are weak, resulting in global sampling. As we move to lower temperatures, the springs become stiffer, focusing the sampling on local regions.

Methods

We can use information coming from experiments (e.g. cross-linking experiments, solid state NMR, EPR), from bioinformatics (e.g. secondary structure predictions[2], homologous proteins[3]), or evolution (e.g. residue-residue contacts predicted from coevolution[4]). These types information share two properties. First, the information can be sparse. For example, an experiment may give us a few residue-residue contacts, but we may know little about the rest of the structure. Second, this information can be noisy. That is, it may contain errors and ambiguities. For example, many of the contacts predicted from sequence coevolution will be wrong. A successful method must be able to deal with data that is sparse and noisy. We have developed an algorithm that can deal with such data in a rigorous statistical mechanical framework.

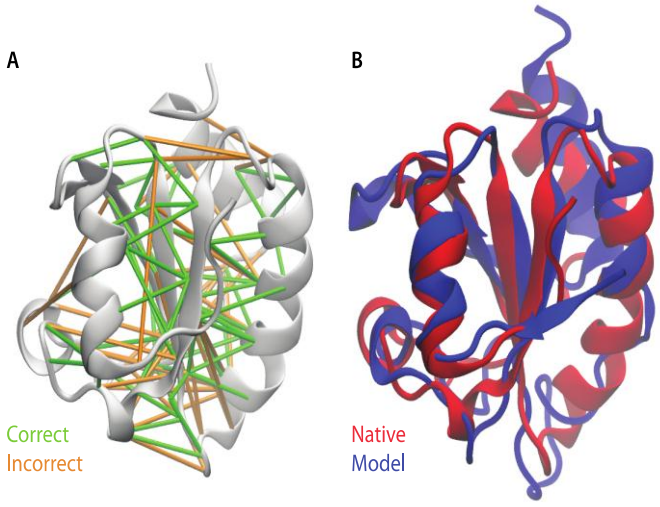


Figure 1: Structure prediction of thioredoxin using contacts predicted from evolution. (A) Distribution of correctly and incorrectly predicted contacts on the native structure. (B) Comparison of the model and the native structures. The Ca RMSD is 3.5 Å.

We use HREMD simulations and a technique we call zipping to efficiently generate structures compatible with the available data. These simulations are run using the GPU accelerated version of Amber, which gives a 100-fold speedup relative to CPUs and allows for much more conformational sampling. The method obeys detailed balance and produces a Boltzmann weighted ensemble, which is important because it means that

we can use populations and free energy to select the correct structures. This method requires far less restraint information per residue (as little as 0.4 restraints/residue) that would be needed to determine the structure by NMR alone (~20 restraints/residue).

Results

Figure 1 shows the predicted structure of thioredoxin, using noisy data from evolution-based contact prediction[5]. Of the 105 contacts predicted, only 68 of them are correct, and trying to enforce all contacts simultaneously would lead to major distortions in the structure. Instead, we assume that the contact predictions are 60% accurate, which we infer based on past performance of the contact prediction algorithm. Using predicted contacts and an assumption of 60% accuracy; our method produces a structure that is 3.5 Å from the crystal structure.

Through out CASP10 we have been relying mostly on bioinformatics tools such as secondary structure prediction and homology modeling to use as restraining information in our simulations. Our methods for ranking structures were mostly based on clustering the ensembles and selecting representative structures from the most populated clusters.

Availability

The main software for the calculations is Amber (http://ambermd.org), and we apply several python wrappers around it to manage and control the restraints that go into Amber at each state.

1. Affentranger, R., Tavernelli, I. & Di Iorio, E. E. (2006). A novel Hamiltonian replica exchange MD protocol to enhance protein conformational space sampling. *Journal of Chemical Theory and Computation* **2**, 217-228.
2. Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology* **292**, 195-202.
3. Remmert, M., Biegert, A., Hauser, A. & Soding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods* **9**, 173-175.
4. Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T. & Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences of the United States of America* **108**, E1293-E1301.
5. Hopf, T. A., Colwell, L. J., Sheridan, R., Rost, B., Sander, C. & Marks, D. S. (2012). Three-Dimensional Structures of Membrane Proteins from Genomic Sequencing. *Cell* **149**, 1607-1621.

Ranking protein structures using free energy as a scale

A. Roy[1] ,J.L.MacCallum[1], A. Perez[1] and K.A. Dill[1]

[1] – *Laufer Center for physical and quantitative Biology, Stony Brook University*

justin.maccallum@me.com

Free energy has a special importance in structural biology. For example, free energy dictates protein folding: small globular proteins have a unique, thermodynamically stable native structure at the global free energy minimum. We use this property for ranking protein structures. However, if the states of interest are very different, various problems like timescale, reaction coordinate, and convergence arises in free energy calculation using direct molecular dynamics simulations or using popular method like umbrella sampling. We further modify and use recently developed confinement free energy method[1-2] to do such calculation in a computationally less expensive and accurate way.

Methods

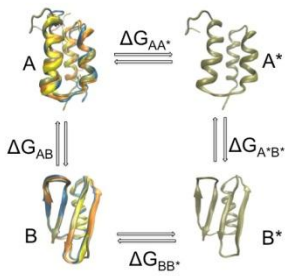


Figure 1: The thermodynamic cycle involving confinement method

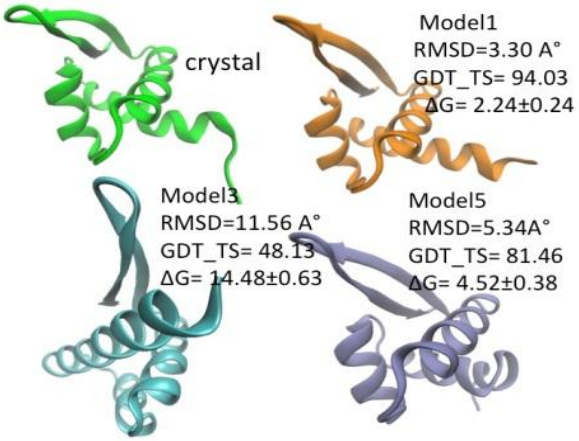
The confinement method uses a thermodynamic cycle. As shown in figure 1, the free energy, ΔG_{AB} between states A and B is calculated in the following way. First, both A and B are converted into a confined state A* and B* by gradually applying larger and larger harmonic restraint on all atoms so that any rotational contribution of the protein is frozen out, and the only remaining motion is vibrational. In order to achieve this, a series of MD calculation are carried out, and the free energy of the confinement, ΔG_{AA^*} and ΔG_{BB^*} are calculated using a numerical approach developed by Tyka et. al.[1] Finally, to close the thermodynamic cycle, the free energy difference between the final restrained /frozen states are calculated using normal mode/quasi-harmonic analysis. The overall free energy of the process is $\Delta G_{AB}=\Delta G_{AA^*} - \Delta G_{BB^*}+\Delta G_{A^*B^*}$. The method does not require a reaction coordinate or transition path and it is computationally less expensive. Moreover, with the advancement of GPU computers this method is fast to compute. In all cases, we use CASP-hosted server predictions as a starting point.

Results

We tested this method with different models of CASP9 prior to applying it to targets of CASP10. In figure 2, we have shown 3 submitted models of Target 559 of CASP 9 along with the calculated free energy values. The group BAKER-ROSETTASERVER submitted all these models. The order of free energies matches the order calculated by RMSD and GDT_TS (Global distance test score) values.

Figure 2: Structures and free energy values for the crystal structure and three submitted models of Target 559 of CASP9. All free energy values are in Kcal/Mol. The calculated RMSD values are backbone RMSD only.

Availability



The main software for the calculations is Amber (http://ambermd.org). We wrote some scripts to manage the whole calculation.

1. Tyka, M., Clarke, R. and Sessions, R. An efficient, path-independent method for free-energy calculation, J. Phys. Chem. B, 110, p. 17212-17220 (2006).

2. Cecchini, M., Krivov, S.V., Spichty, M., Karplus, M. Calculation of free-energy differences by confinement simulations. Application to peptide conformers. J. Phys. Chem. B 113, p. 9728-9740 (2009).

Protein Modeling System by global optimization

Keehyoung Joo[1,2], Juyong Lee[1], Sangjin Sim[1], Kiho Lee[1], Seungryong Heo[1], Sun Young Lee[1], In-Ho Lee[1,3], Sung Jong Lee[1,4], and Jooyoung Lee[1,2*]

[1]Center for In-Silico Protein Science, Korea Institute for Advanced Study, 130-722, Korea, [2]Center for Advanced Computation, Korea Institute for Advanced Study, 130-722, Korea, [3]Korea Research Institute of Standards and Science (KRISS), 305-600, Korea, [4]Department of Physics, University of Suwon, Hwaseong-Si, 445-743, Korea
*jlee@kias.re.kr

We have developed a Protein Modeling System (PMS) for sever prediction based on global optimization of energy functions and quality assessment for protein 3D models, and we have applied it to all CASP10 targets. PMS method adds additional new features to the old *gws*/LEE [1-3] method which was used in CASP9. Here, we focused on developing new energy functions including new restraint terms and physical energy terms to build protein 3D models. In addition, we developed a new quality assessment method to select template candidates as well as multiple alignments.

Energy function to build protein 3D models:
For protein 3D modeling, we developed a new Lorentzian-type energy term for structural restraints instead of using gaussian type or spline functions used in MODELLER. In order to obtain the sigma values for the width and depth of the Lorentzian functions (which control the strength of individual distance restraint), we employed a machine learning method, a random forest algorithm to predict the sigma values, where the input features are based on the sequence-template alignment and environmental features including the profile similarity, gap features, secondary structure consensus, and solvent accessibility consensus. For loop regions for which no restraints are available from templates, we combined physical energy terms including dynamic fragment assembly (DFA) energy (which were originally developed for *ab-initio* protein structure modeling[4]) together with DFIRE statistical potential energy, hydrogen bonding term, and GOAP terms[5]. In order to optimize the energy function, we utilized conformational space annealing (CSA), a powerful global optimization and efficient conformational search method. The energy parameters and weights for energy terms were trained using a subset of CASP9 targets.

Model quality assessment to select templates and alignments:
For selecting templates and alignments, we developed a new quality assessment (QA) method using the random forest algorithm. For feature vectors, we used the energy values of individual energy terms described above evaluated from predicted 3D models and consensus features between secondary structure & solvent accessibility of the 3D models and the corresponding predicted values. In the fold recognition step, the QA method was used to re-rank template candidates generated by FOLDFINDER, an in-house sequence-template alignment method. Multiple sequence-template alignment (MSA) for each combination among the query sequence and its templates was carried out by using MSACSA[2], a multiple sequence alignment method by global optimization. Then, the QA method was used to select the best MSA candidate by assessing each 3D model generated from templates and its multiple alignments.

With these major changes in the protocol, we performed model optimizations and side-chain re-modeling by successively applying the conformational space annealing in PMS protocol. We applied three human prediction methods. One human prediction method of LEE applied the same protocol of PMS and, in addition, considered additional templates from FOLDFINDER. Another human prediction method of LEEMO uses multi-objective CSA (MOCSA[6]) optimization instead of single-objective CSA in the model building step. And the other human prediction method of LEEcon is consensus method using SERVER prediction models as templates, which are identified in the largest cluster.

Acknowledgements: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 20120001222 and 2009-0090085). We thank Korea Institute for Advanced Study for providing computing resources (KIAS center for Advanced Computation Linux Cluster) for this work.

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, J., Sasaki, T. N., Seok, C., & Lee, J. (2011) De novo protein structure prediction by dynamic fragment assembly and conformational space annealing, *Proteins*, **79**, 2403-2417.
5. Zhou, H., & Skolnick, J. (2011) GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction, *Biophys. J.*, **101**(8), 2043-2052.
6. Sim, S., Lee, J., & Lee, J., (2012) MOCSA: multi-objective optimization by conformational space annealing, arXiv:1209.0549.

MATRIX

Methodology for Accurate Template Recognition for predIcting X [=proteins] server

G. Chopra[1, 2], H. Tiang[1] and R. Samudrala[1]

[1] - *Dept. of Microbiology University of Washington USA,* [2] - *Dept. of Structural Biology Stanford University USA*
ram@compbio.washington.edu

Template-based modeling has been the most successful method at CASP till date. Selection of templates remains a difficult task; it is not always possible to identify the best template from the PDB due to the limitations of the alignment methods. Using a large number of threading programs has become a routine to select templates with best alignments. However, in many cases even though the best template is identified by the threading programs, it is ranked lower based on the threading Z-scores and is buried deep within many templates. MATRIX is an automated protein 3D structure prediction server, which proposes a solution to pick such missed templates using binding site information. Specifically, we use the premise that the functional sites are conserved across multiple templates identified by the threading programs and combine restraints from multiple threading alignments for structure prediction.

Methods

We address the problem of template recognition from a large set of alignments from multiple threading programs, specifically LOMETS[1] and HHSearch[2] programs. All the templates were collected and the binding site similarity among them were compared using the COFACTOR methodology[3; 4]. The final selection of templates is based on the threading score and alignment coverage cutoff values in addition to the templates identified by high local similarity binding score (BS-score) in COFACTOR. We combined the restraints from multiple threading alignments of all the identified templates and used MODELLER program[5] to quickly generate the model. Finally, we refined the top five models generated by MODELLER using the consistent refinement protocol implemented by the KoBaMIN refinement protocol[6; 7] (http://csb.stanford.edu/kobamin/) for submission.

Availability

MATRIX is available as a web sever at http://cando.compbio.washington.edu/casp/matrix.

1. Wu S. & Zhang, Y. (2007) LOMETS: A local meta-threading-server for protein structure prediction. *Nucleic Acids Res* **35**, 3375-3382.
2. Soding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**, 951-60
3. Roy, A., Yang, J. and Zhang, Y. (2012) COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Res*. **40**, W471-477.
4. Roy, A. and Zhang, Y. (2012) Recognizing protein-ligand binding sites by global structural alignment and local geometry refinement. *Structure*. **20**, 987-997.
5. Fiser A. & Sali A. (2003) Modeller: generation and refinement of homology based protein structure models. *Methods Enzymol*. **374**, 461–491.
6. Chopra,G., Kalisman,N. & Levitt,M. (2010). Consistent refinement of submitted models at CASP using a knowledge-based potential. *Proteins*. **78**, 2668-2678.

7. Rodrigues, J. P. G. L. M., Levitt, M. & Chopra, G. (2012). KoBaMIN: a knowledge-based minimization web server for protein structure refinement. *Nucleic Acids Res.* **40**, W323-W328.

Manual Predictions of the Tertiary Structures of Proteins and their Homo-Multimeric States

M.T. Buenavista[1,2,3], D.B. Roche[4,5,6], E.J. Fox[1] and L.J. McGuffin[1]

[1] - *School of Biological Sciences, University of Reading, Reading, UK,* [2] - *Biocomputing Section, Medical Research Council Harwell, Harwell Oxford, Oxfordshire, UK,* [3] - *Beamline B23, Diamond Light Source, Didcot, UK.* , [4] - *Commissariat à l'énergie atomique et aux énergies alternatives, Institut de Génomique, Genoscope, 2 rue Gaston Crémieux, 91057, Evry, France,* [5] - *Centre National de la Recherche Scientifique, UMR 8030, 2 rue Gaston Crémieux, 91057, Evry, France,* [6] - *Université d'Evry-Val-d'Essonne, Boulevard François Mitterrand, 91025, Evry, France*

l.j.mcguffin@reading.ac.uk

For our manual predictions we made use of the component methods that we developed for the IntFOLD2 and ModFOLD4 servers[1,2] (see our server abstracts for more detail). However, we also made use of all of the provided 3D server models, we heavily relied on our Quality Assessment predictions obtained from ModFOLDclust2[3] and we used a considerable amount of manual intervention as we were developing our prototype homo-multimeric prediction protocol, detailed below.

Methods

Tertiary Structure Predictions: For the tertiary structure (TS) category, our manual predictions were made using ModFOLDclust2[3] 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.

Homo-Multimer Predictions: Our multimeric prediction protocol made use of the ModFOLDclust2 selected 3D server models and the lists of templates generated by our IntFOLD2 server[1], which has recently been updated to include multi-template modeling[5] (see our IntFOLD2 abstract for details).

The homo-multimeric modeling for each target involved: 1. Ranking 3D model-template alignments according to their combined mean coverage and TM-scores using TM-align[6] 2. Extracting the multimeric state information for each template from PISA[7] 3. Filtering templates based on the PISA multimeric assembly annotation and on set TM-score thresholds. 4. Building N-meric model assemblies based on selected N-mer PISA templates, using TM-align and PyMOL (http://www.pymol.org) to orientate the model subunits. 5. Implementing further screening criteria that included parameters such as MMalign[8] scores between multimer

templates and multimer models, alignment lengths and alignment coverage and 6. Rating interface quality by taking into account the predicted B-factors of interface residues, clashes or overlaps and the distance of contacting units or sub-units, in an attempt to select the most appropriate modeled assembly.

Availability

Our homo- multimeric prediction algorithm is currently being automated and will shortly be integrated with the IntFOLD2 server:
http://www.reading.ac.uk/bioinf/IntFOLD/IntFOLD2_form.html.
The ModFOLDclust2 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. Roche,D.B., Buenavista,M.T., Tetchner,S.J. & McGuffin,L.J. (2011). The IntFOLD server: an integrated web resource for protein fold recognition, 3D model quality assessment, intrinsic disorder prediction, domain prediction and ligand binding site prediction. *Nucleic Acids Res.* **39**, W171-6.
3. 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.
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. Buenavista, M.T., Roche, D.B., & McGuffin, L J. (2012) Improvement of 3D protein models using multiple templates guided by single-template model quality assessment. *Bioinformatics*. **28**, 1851-1857.
6. Zhang, Y. and Skolnick, J. (2005) TM-align: A protein structure alignment algorithm based on TM-score, *Nucleic Acids Res*. **33**, 2302-9.
7. Krissinel, E. and Henrick. K. (2007). Inference of macromolecular assemblies from crystalline state. J. *Mol. Biol.* **372**, 774-97.
8. Mukherjee, S. and Zhang, Y. (2009) MM-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming. *Nucleic Acids Res*. **37**, e83.

Manual Ligand Binding Site Residue Predictions

D.B. Roche[1,2,3], M.T. Buenavista[4,5,6] and L.J. McGuffin[4]

[1] - Commissariat à l'énergie atomique et aux énergies alternatives, Institut de Génomique, Genoscope, 2 rue Gaston Crémieux, 91057, Evry, France, [2] - Centre National de la Recherche Scientifique, UMR 8030, 2 rue Gaston Crémieux, 91057, Evry, France, 3 - Université d'Evry-Val-d'Essonne, Boulevard François Mitterrand, 91025, Evry, France, 4 - School of Biological Sciences, University of Reading, Reading, UK, [5] - Biocomputing Section, Medical Research Council Harwell, Harwell Oxford, Oxfordshire, UK and [6] -Beamline B23, Diamond Light Source, Didcot, UK.

l.j.mcguffin@reading.ac.uk

Our manual ligand binding site residue predictions were for the most part automated, using the output from our IntFOLD2 server[1]. We also made use of the 3D server models, heavily relying on QA results from ModFOLDclust2[2] and FunFOLDQA[3], along with some manual intervention in an attempt to add value to our automated FunFOLD[4] predictions.

Methods

We attempted to add value to our automated binding site residue predictions using the standalone version of FunFOLD[4], better server 3D starting models, QA information from ModFOLDclust2[2] and the standalone version of our new ligand binding site quality prediction method FunFOLDQA[3].

Firstly, for the server only targets, the FunFOLD[4] server results were visually inspected and included in the prediction based on the following criteria: 1. The global quality score for the starting model was acceptable; 2. Residues were in contact with more than two well superposed ligands.

Secondly, for manual targets ModFOLDclust2[2] was utilized to rank the server models. The standalone version of FunFOLD[4], was subsequently used to predict ligand binding site residues for the top 10 server models. The 10 resultant FunFOLD[4] predictions were ranked using the standalone version of our ligand binding site quality assessment tool FunFOLDQA[3], which produces predictive MCC[5] and BDT[6] scores. The FunFOLD[4] prediction with the highest predicted MCC and BDT scores was submitted as our manual prediction, if after visual inspection the following criteria held true: 1. The global quality score for the start model was acceptable; 2. The residues were in contact with more than two well superposed ligands; 3. The model-to-template superpositions were good.

Results

Preliminary results indicate some improvements for the manual FN predictions over the server FN predictions. Taking target T0726 as an example, the server prediction achieves an MCC score of 0.773 and a BDT score of 0.612, whereas the manual prediction achieves an MCC and BDT score of 1.0.

Availability

The IntFOLD2 server with graphical output is available at:

http://www.reading.ac.uk/bioinf/IntFOLD/IntFOLD2_form.html
The BDT, FunFOLD, FunFOLDQA and ModFOLDclust2 software can be downloaded from:
http://www.reading.ac.uk/bioinf/downloads/

1. Roche,D.B., Buenavista,M.T., Tetchner,S.J. & McGuffin,L.J. (2011). The IntFOLD server: an integrated web resource for protein fold recognition, 3D model quality assessment, intrinsic disorder prediction, domain prediction and ligand binding site prediction. *Nucleic Acids Res.* **39**, W171-6.
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**, 182-8.
3. Roche,D.B., Buenavista,M.T. & McGuffin,L.J. (2012). FunFOLDQA: a quality assessment tool for protein-ligand binding site residue predictions. *PLoS One* **7**, e38219.
4. Roche,D.B., Tetchner,S.J. & McGuffin,L.J. (2011). FunFOLD: an improved automated method for the prediction of ligand binding residues using 3D models of proteins. *BMC Bioinformatics* **12**, 160.
5. Matthews,B.W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* **405**, 442-51.
6. Roche,D.B., Tetchner,S.J. & McGuffin,L.J. (2010). The binding site distance test score: a robust method for the assessment of predicted protein binding sites. *Bioinformatics* **26**, 2920-1.

BCL::Fold - De novo prediction of complex and large protein topologies by assembly of secondary structure elements

S. Heinze[a], M. Karakaş, N. Woetzel, R. Staritzbichler, N. Alexander, B. Weiner, D. Putnam, J. Meiler[a]

[a]*Chemistry Department, Vanderbilt University*
Sten.Heinze@vanderbilt.edu

Computational *de novo* protein structure prediction is limited to small proteins of simple topology. BCL::Fold introduces an algorithm for protein folding with a novel approach of assembling secondary structure elements (SSEs) in three-dimensional space. Our approach seeks to overcome size and complexity limits of previous methods by discontinuing the amino acid chain in the folding simulation. This facilitates the sampling of non-local contacts. By excluding the loop regions, we focus the sampling to the relative arrangement of rather rigid SSEs thus limiting the search space. We leverage established protocols for construction of loop regions and side chains to yield complete protein models. Decoupling the placement of SSEs from the construction of loop regions relies on the excessively tested hypothesis[1,2] that accurate placement of SSEs will allow for construction of loop regions and side chains.

Methods

The Monte Carlo Metropolis-based algorithm uses simulated annealing and SSE-based moves to alter the protein models. It optimizes a knowledge-based potential that consists of twelve individual terms: amino acid pair distance clash, amino acid pair distance, amino acid solvation, SSE pair clash, SSE pair packing, β-strand pairing, loop length, strictly enforcing loop closure, radius of gyration, SSE prediction for JUFO, SSE prediction for PSIPRED, and lastly contact order. All knowledge based potentials have been derived from a databank that contained 3,409 high resolution x-ray crystallography protein structures compiled using the PISCES server[3].

We use the secondary structure prediction programs JUFO[4,5] and PSIPRED[6] to create a comprehensive pool of predicted SSEs. To avoid incorrectly predicted secondary structure we implement two strategies: a) multiple copies of one SSE of different lengths and types are collected; b) the lengths of SSEs are adjusted during the folding simulation in order to optimize protein secondary and tertiary structure prediction[4].

The minimization process contains two stages. The “assembly” stage consists of large amplitude translation or rotations and addition or removal of SSEs. The “refinement” stage focuses on small amplitude moves that maintain the current topology but optimize interactions.

Once the SSE pool is input, the algorithm initializes both the energy functions and the move sets for assembly and refinement stages. A starting model for the minimization is created by inserting a randomly selected SSE from the pool into an empty model. The starting model is passed to the minimizer which executes assembly and refinement minimization. The assembly stage terminates after 5000 steps or after 1000 consecutive steps without score improvement. The refinement stage terminates after 2000 steps or after 400 consecutive steps without score improvement. In general, a move can result in one of four outcomes: “improved” in score, “accepted” through Metropolis criterion, “rejected” as score worsened, or “skipped” if SSE elements required for the move are not present in the model. The temperature is adjusted dynamically based on the ratio of accepted steps.

For each CASP10 target 12,000 models were generated, the top 50% by BCL score were

selected for clustering analysis. The best scoring models as well as the best scoring models in each of the large clusters underwent loop construction and side chain packing using ROSETTA. Up to five models for submission were selected from these full atom models.

Availability

The described method will be made available in two ways. A web-accessible service will be provided to test the method and obtain a limited number of model predictions. Executable programs for different system environments and additional protocol information will be made available for download under academic and commercial licenses within the BCL::Commons set of applications. Both will be obtainable at http://www.meilerlab.org.

1.Kuhlman, B. & Baker, D. Native protein sequences are close to optimal for their structures. *Proc Natl Acad Sci U S A* **97**, 10383–10388 (2000).

2.Bradley, P., Misura, K. M. S. & Baker, D. Toward high-resolution de novo structure prediction for small proteins. *Science* **309**, 1868–1871 (2005).

3.Wang, G. & Dunbrack, R. L., Jr PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Res.* **33**, W94–98 (2005).

4.Meiler, J. & Baker, D. Coupled prediction of protein secondary and tertiary structure. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 12105–12110 (2003).

5.Jens Meiler, M. M. Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. **7**, 360–369 (2001).

6.Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195–202 (1999).

BCL::SAXS – Small Angle X-Ray Scattering Profiles to Promote Protein Folding

D. Putnam[ab], B. Weinera[c], N. Woetzel, S. Heinze[c], E. Lowea[c], J. Meiler[a,b,c]

a- Center for Structural Biology, Vanderbilt University
b - Biomedical Informatics Department, Vanderbilt University
c - Chemistry Department, Vanderbilt University

daniel.k.putnam@vanderbilt.edu

Small Angle X-Ray Scattering (SAXS) is often used for low resolution structural characterization of proteins that evade other experimental techniques, such as x-ray crystallography and nuclear magnetic resonance (NMR).[1] Here, we introduce BCL::SAXS – an algorithm designed to replicate SAXS curves from rigid idealized protein models. We first show our derivation of BCL::SAXS and compare our results with CRYSOL[2] for 1) complete protein models, 2) models without side chain coordinates, and 3) models without side chains and loop regions. We evaluate the ability to identify a correct protein topology from a set of 455 proteins from the PISCES dataset with 20% identify cutoff, 1.6 Å resolution cutoff, and 025 R-factor cutoff.[3] The SAXS score was 99% accurate in identifying the correct protein topology from a large set of different protein topologies. Further, we evaluate the effect of using the SAXS score as a weighted term in the knowledge-based energy function of BCL::Fold for seven soluble protein examples. BCL::SAXS increased the fraction of correctly folded models for proteins with extended topologies, but did not increase the fractions of correctly folded models for globular proteins.

Methods

To accurately determine the SAXS profile from the atomic coordinates of full atom protein models we utilized the Debye formula for atomic scatterers and associated equations to calculate the form factors.[4-6] We used GPU acceleration to parallelize the pair wise computation of the Debye formula. Once scattering profiles were generated from rigid body protein models we compared the scattering profile computed by BCL::SAXS with the scattering profile computed by CRYSOL. To compare the profiles, we used a cubic spline function to compute the first derivative of each curve. A $\chi 2$ measure was used to quantify the difference between the derivatives of the two scattering curves. This measure is the computed SAXS score. We scored a random subset of 455 proteins from the PISCES data set with each other to verify the saxs score could distinguish protein topologies.

To approximate the side chain regions of a given amino acid, the form factors for the atoms with missing side chain coordinates were added to the Cβ position of the respective amino acid. The loop regions were approximated by removing atomic coordinate data between secondary structure elements (SSEs) and computing a parabolic path from the c-terminus of the first SSE to the n-terminus of the second SSE. The amino acid residues in the loop regions were placed at points along the path.

The BCL::SAXS score was added to the minimization process in BCL::FOLD. During the first two rounds of assembly the weight of the SAXS score was set to zero. The score was weighted successfully higher in stages three, four and five. During the refinement stage the saxs score was weighted lower. Each round of the assembly stage terminates after 2000 steps or after

500 consecutive steps without score improvement. The refinement stage terminates after 4000 steps or after 500 consecutive steps without score improvement.

Availability

The described method will be made available in two ways. A web-accessible service will be provided to test the method and obtain a limited number of model predictions. Executable programs for different system environments and additional protocol information will be made available for download under academic and commercial licenses within the BCL::Commons set of applications. Both will be obtainable at http://www.meilerlab.org.

1. C. D. Putnam, M. Hammel, G. L. Hura, and J. A. Tainer, "X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution," Q Rev Biophys, vol. 40, pp. 191-285, Aug 2007.
2. D. Svergun, C. Barberato, and M. H. J. Koch, "CRYSOL - A program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates," Journal of Applied Crystallography, vol. 28, pp. 768-773, Dec 1 1995.
3. G. Wang and R. L. Dunbrack, Jr., "PISCES: recent improvements to a PDB sequence culling server," Nucleic Acids Res, vol. 33, pp. W94-8, Jul 1 2005.
4. P. Debye, "Zerstreuung von Röntgenstrahlen," Annalen der Physik, vol. 351, pp. 809-823, 1915.
5. D. Schneidman-Duhovny, M. Hammel, and A. Sali, "FoXS: a web server for rapid computation and fitting of SAXS profiles," Nucleic Acids Res, vol. 38, pp. W540-4, Jul 2010.
6. K. Stovgaard, C. Andreetta, J. Ferkinghoff-Borg, and T. Hamelryck, "Calculation of accurate small angle X-ray scattering curves from coarse-grained protein models," BMC Bioinformatics, vol. 11, p. 429, 2010.

Prediction of protein disordered regions based on meta approach

Takashi Ishida[1]

[1] – *Department of Computer Science, Tokyo Institute of Technology*
t.ishida@bi.cs.titech.ac.jp

"metaprdos2" is an automated protein disordered region prediction server based on meta prediction approach. The prediction system is basically same as metaPrDOS server[1] but used five independent predictors and modified input vector for second prediction.

Methods

The prediction comprises two main steps. In the first step, an input sequence is submitted to each disorder predictor, and prediction results from all predictors are collected. We used five predictors: PrDOS2, DISOPRED2, DISPROT (VSL2P), DISpro, and POODLE-S. Each predictor will perform its own prediction for each residue, and the result is obtained as a scaled value. In the second step, the meta predictor integrates the prediction results and determines the disorder tendency for each residue. The input vector of meta prediction includes not only the output of each component predictor but also the number of homologues sequences in NCBI nr and pdbaa to a target sequence. Because, some component predictor shows lower performance without homologues sequence information as shown in evaluation of disorder prediction in CASP7[2]. Especially, the performance of PrDOS2 highly depends on the homologues in the PDB. We adopted the support vector machine (SVM) as the prediction algorithm. Finally, the decision value of a SVM is scaled from 0.0 to 1.0, and it is returned as a prediction result.

1. Ishida, T and Kinoshita, K. (2008). Prediction of disordered regions in proteins based on the meta approach, *Bioinformatics*, **24**, 1344-1348
2. Bordoli, L. *et al.*, (2009) Assessment of disorder predictions in CASP7. *Proteins*, **69**, 129-136.

Automated 3D Model Quality Assessment using the ModFOLD4 Server

L.J. McGuffin[1], D.B. Roche[2,3,4] and M.T. Buenavista[1,5,6]

[1] - *School of Biological Sciences, University of Reading, Reading, UK,* [2] - *Commissariat à l'énergie atomique et aux énergies alternatives, Institut de Génomique, Genoscope, 2 rue Gaston Crémieux, 91057, Evry, France,* [3] - *Centre National de la Recherche Scientifique, UMR 8030, 2 rue Gaston Crémieux, 91057, Evry, France,* [4] - *Université d'Evry-Val-d'Essonne, Boulevard François Mitterrand, 91025, Evry, France,* [5] - *Biocomputing Section, Medical Research Council Harwell, Harwell Oxford, Oxfordshire, UK and* [6] - *Beamline B23, Diamond Light Source, Didcot, UK.*

l.j.mcguffin@reading.ac.uk

The ModFOLD4 server is the latest version of our popular resource for the Quality Assessment (QA) of 3D models of proteins [1,2].

Methods

The new version of the ModFOLD server is capable of working in quasi-single model mode or in multiple-model/clustering mode. The first stage of the algorithm generates ~84 Tertiary Structure (TS) models using the novel multi-template approach[3] which forms the basis of the IntFOLD2 server (see our IntFOLD2 abstract for more details).

In the default server mode (ModFOLD4), a straightforward clustering approach was used whereby all submitted models were pooled together with the IntFOLD2 TS models and clustered using ModFOLDclust2[4] (see our ModFOLDclust2 abstract for further details). In addition, we decided to include a forced single-model mode version of the server (ModFOLD4_single) in order to simulate the effect of users submitting one model at a time. Therefore, in quasi-single model server mode each submitted model was compared in isolation against the pool of IntFOLD2 models using a global and local scoring approach similar to that used by ModFOLDclust2.

Availability

The ModFOLD server version 4.0 is available at the following URL:
http://www.reading.ac.uk/bioinf/ModFOLD/ModFOLD_form_4_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. (2009) Prediction of global and local model quality in CASP8 using the ModFOLD server. *Proteins*. **77**, 185-190.
3. Buenavista, M. T., Roche, D. B. & McGuffin, L. J. (2012) Improvement of 3D protein models using multiple templates guided by single-template model quality assessment. *Bioinformatics*. **28**, 1851-1857.
4. 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.

Automated 3D Model Quality Assessment using ModFOLDclust2

L.J. McGuffin[1], D.B. Roche[2,3,4]

[1] - *School of Biological Sciences, University of Reading, Reading, UK,* [2] - *Commissariat à l'énergie atomique et aux énergies alternatives, Institut de Génomique, Genoscope, 2 rue Gaston Crémieux, 91057, Evry, France,* [3] - *Centre National de la Recherche Scientifique, UMR 8030, 2 rue Gaston Crémieux, 91057, Evry, France,* [4] - *Université d'Evry-Val-d'Essonne, Boulevard François Mitterrand, 91025, Evry, France*

l.j.mcguffin@reading.ac.uk

The ModFOLDclust2 method[1] is a leading automatic clustering based approach for both local and global 3D model quality assessment[2].

Methods

The ModFOLDclust2 method tested at CASP10 was identical to that tested in CASP9. The ModFOLDclust2 method was originally 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 ModFOLDclustQ method and the original ModFOLDclust method[3,4]. ModFOLDclustQ is similar to our previous ModFOLDclust method, 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 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.

Availability

ModFOLDclust2 is provided as a program option via the ModFOLD server version 3.0:
http://www.reading.ac.uk/bioinf/ModFOLD/ModFOLD_form_3_0.html
The ModFOLDclust2 software is also available to download as a standalone program via:
http://www.reading.ac.uk/bioinf/downloads/

1. 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.
2. Kryshtafovych A, Fidelis K, Tramontano A.(2011) Evaluation of model quality predictions in CASP9. *Proteins*.**79**, Suppl 10:91-106.
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.

MQAPmulti2 and MQAPsingle2: toward the estimation of model quality when not only many models are available

M. Pawlowski[1] and A.Kloczkowski[1,2]

[1]-*Battelle Center for Mathematical Medicine,The Research Institute at Nationwide Children's Hospital ,Columbus,OH, USA*
[2]-*The Ohio State University, Columbus, OH, USA*
Marcin.Pawlowski@nationwidechildrens.org

At the CASP9 meeting several quality assessment issues have been raised. One of them was linked to performance of clustering MQAPs that are advantaged in comparison to single-model methods on sets of hundreds of models. In such cases the high performance of clustering MQAP is caused mainly by the fact that by clustering, these MQAPs can easy identify models of incorrect fold, however their performance drops significantly once the ranking of the top quality models is considered. In real-life scenario, the user usually wants to predict the quality for a few models that were produced by best modeling servers.

Methods

In this work, we introduce a method for the estimation of quality of structure models. Not only does the MQAP perform well when scoring hundreds of alternative models, but also it can be applied when only a few models (~20) are available. To do so, we optimized MQAPmulti (a program developed by the first author of this abstract, but in different laboratory) to perform better when 20 or 150 models are available.

Likewise MQAPmulti, the MQAPmulti2 prediction is based on the three following components: 1) TrueMQAP_componet – scoring functions that is based on statistical and agreement potentials; 2) CLUST_component, which clusters models on the base of GDT_TS[1] and SQ_score (our modification of Q-score[2] that works by estimating the structural relatedness between two protein structures based on comparison of intramolecular distances); 3) CORR_component, a correlation based method that combines predictions of the TrueMQAP_componet with pair-wise models comparisons measured by GDT_TS and SQ_score. Finally, all of these components are used to predict the global quality of a model. To do so, on the base of the number of models to score, the program chooses one of 3 regression models that describe the relationship between initial parameters and the global quality. These three regression models were created for following numbers of input models: 20, 150, 300 or more.

MQAPsingle2, that is a variant of the MQAPmulti2 program, operates as a quasi-single model MQAP. This method applies MQAPmulti2 algorithm, but a model is not compared to the input models, but to models generated by GeneSilico fold prediction metaserver[3].

Results

MQAPmulti2 was trained and tested for *CASP7th, 8th and 9th* models dataset, 10-fold cross validation procedure was applied to do so. The value of Pearson's correlation coefficient between MQAPmulti2 global score and the GDT_TS of models is 0.712, 0.819 and 0.917.

Availability

The MQAPmulti2 and MQAPsingle2 can be executed as standalone programs.

1. Zemla, A. (2003). LGA—a method for finding 3D similarities in protein structures. Nucleic Acids Res. 31, 3370–3374.
2. Goldstein,R.A. et al. (1992) Optimal protein-folding codes from spin-glass theory. Proc.Natl Acad. Sci. USA. 89, 4918–4922.
3. Kurowski MA, Bujnicki JM. (2003) GeneSilico protein structure prediction meta-server. Nucleic Acids Res. 31(13):3305-3307.

Evolutionary Algorithms for Protein Model Refinement

S. Nguyen[1], Y. Chen[1], W. Xiong[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, Columba, 65211, USA
spnf2f@mail.missouri.edu, shangy@ missouri.edu

Given a set of predicted protein models from multiple sources or methods, an important task is to utilize their diverse information to generate better models[1]. Similarly, if the best model in the set is known, it could be improved by using information in other models. In this work, several evolutionary algorithms have been developed to address these tasks.

Methods

The main framework of the evolutionary algorithms consists of three phases: Model Selection, Crossover, and New Model Selection. The basic idea is to select a few models from a population of models, such as 5, give them to Modeller to generate several new models, and put the good ones back into the model population.

In the Model Selection phase, two different strategies are used for selection: random selection and seeded selection. In random selection, a few models are randomly selected from the model pool to form the input set to Modeller. In seeded selection, models in the model pool are first scored using our quality assessment (QA) method MUFOLD-QA[2]. Then, a seed model is selected with a probability in proportion to its score. Next, several models similar to the seed model, i.e., within a certain GDT_TS value, are randomly selected. These models, together with the seed model, form the input set to Modeller. In the evolutionary algorithms, these two strategies are alternated in every other generation to achieve a balance of exploitation around a specific configuration and exploration across a broad range of configurations in the search space.

In the Crossover phase, selected models from the previous phase are fed into Modeller, a program producing homology models of protein tertiary structures from a given set of models, to generate some new models. Empirical results show that Modeller could improve input models slightly, although the output models are generally similar to one or several input models.

In the New Model Selection phase, good new models generated by Modeller are selected and added into the model population. The new models are first scored by our QA method, and then compared with existing models. If some of the new models are better than some existing models, they are added into the model population to replace the worse ones.

These three phases are iterated for a number of generations. The evolutionary algorithms terminate either when the maximum number of generations is reached or the models become too similar and the whole population converges. Through the evolutionary process, the model pool usually becomes better over time. In the end, our QA method is used to evaluate the final pool of models and pick up the best models as the final result.

The evolutionary process with slight variations can be applied to either model refinement where the best model in a pool is known or human prediction where the best model is unknown.

Results

In CASP10, the evolutionary algorithms were implemented in MUFOLD2 server for both human prediction and refinement tasks. For human prediction, the server prediction pool was first evaluated using our QA method and the best 200 models formed the initial population for the

evolutionary algorithm. The difference in refinement is that since the best model in the server prediction pool is given, it was used as the seed when the input model set to Modeller was generated in every other generation. Preliminary experimental results of these evolutionary algorithms on CASP9 targets show small improvement over initial models. More extensive experiments will be conducted on CASP9 and CASP8 data. In addition, their performances on CASP10 data will be evaluated after the native structures of the CASP10 targets are released.

1. Baker D, Sali A. (2001). Protein structure prediction and structural genomics. *Science*, 294:93-96.
2. Wang Q, Shang Y, Xu D. (2011). Improving a Consensus Approach for Protein Structure Selection by Removing Redundancy. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(6):1708-1715.

Hybrid Methods for Protein Model Quality Assessment

S. Nguyen[1], Y. Chen[1], W. Xiong[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, Columba, 65211, USA*
spnf2f@mail.missouri.edu, shangy@ missouri.edu

In protein structure predictions, assessing the quality of predicted models is very important. There are two major approaches for protein model quality assessment (QA): single-model scoring functions and structure-based consensus methods. Each approach has its strength and weakness and performs differently under different circumstances. In CASP7 though CASP9, consensus methods have shown a clear advantage on CASP QA data sets[1]. To address the weakness of consensus methods, in CASP10, two new types of QA tasks, one for 20-models and one for 150-models, were designed. In this work, several hybrid algorithms combining single-model scoring functions and consensus methods were developed, aimed at achieving better performance on the new QA tasks than the individual methods.

Methods

Although consensus QA methods have performed very well on CASP targets with more than 300 server predicted models, they have difficulty on a small number of models or on a set of similar models. On the other hand, the QA performance of single-model scoring functions is independent of the number of models. Combining them could lead to a more robust algorithm.

Our hybrid algorithms are based on three single-model scoring functions, Opus-CA, dDfire and CalRW, and the naive consensus method. The three scoring functions are selected due to their speed and reliability. Since the ranges of these scores are very different, given a set of models, their values are normalized into z-scores with mean 0 and standard deviation 1. Then, the z-scores are normalized to the range of [0, 1]. For each model *i*, the average of the three z-scores from the three scoring functions is its value Z_i. Weight W_z for a set of models is the average Z_i values of all models in the set.

The naive consensus value (C_i) of a model based on a reference model set is the average GDT_TS values of the model against each model in the reference set. Weight W_c for a set of models is the average C_i values of all models in the set.

In MUFOLD-HQA, the hybrid algorithm for the 20-model QA task is as follows:

(1) Generate a set of prediction models using the MUFOLD server[2].

(2) Compute the Z_i values of the MUFOLD models; divide their Z_i values into 20 equal-size bins.

(3) Randomly select one model from each of the 20 bins; combine them with the initial 20 QA models to form a combined set of 40 models.

(4) Re-compute the Z_i values of the 40 models as a set and then compute the weight W_z.

(5) Compute the consensus values C_i for each QA model based on the 40-models reference set. Then compute weight W_c.

(6) Compute final QA score of each QA model:

The 150-model QA is similar to the 20-model QA with two major differences. The first difference is that MUFOLD-WQA[1] instead of the naive consensus method is used. The second

difference is that MUFOLD models are not used. Therefore, given 150 QA models, the algorithm computes the normalized Z_i values based on the 150 models. Then, it computes the consensus values C_i using MUFOLD-WQA. Finally, the QA score of each QA model is a weighted sum of the two, as in Step (6).

The MUFOLD-QA server uses consensus methods only. For the 20-models QA task, a set of template models is generated by the MUFOLD server[2] is used as the reference set for computing the naïve consensus scores. For the 150-models QA task, the MUFOLD-WQA method is used.

Results

In CASP10, MUFOLD-HQA employed hybrid algorithms combining scoring functions and consensus methods, whereas MUFOLD-QA simply used consensus methods. Their performance will be evaluated after the native structures of the CASP10 targets are released.

Availability

1. Q. Wang, K. Vantasin, D. Xu, and Y. Shang, "MUFOLD-WQA: A New Selective Consensus Method for Quality Assessment in Protein Structure Prediction," *Proteins*, 79(S10):185-195, 2011.
2. 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*, 78(5):1137-1152, 2009.

Protein Tertiary Structure Prediction Guided by Multi-layer Quality Evaluations

Jingfen Zhang[1,3], Zhiquan He[1,3], Jiong Zhang[2], Son Nguyen[1], Ioan Kosztin[2], Yi Shang[1], and Dong Xu[1,3]

[1]Department of Computer Science, [2]Department of Physics and Astronomy, [3]Christopher S. Bond Life Sciences Center, University of Missouri, Columba, 65211, USA

zhangjingf@missouri.edu

We have developed a system, MUFOLD[1], to predict tertiary structure from protein sequence, where a multi-layer evaluation approach is applied to guide the model generation and improve the model quality iteratively.

Methods

The system includes three phases: Template Selection, Model Generation, and Model Selection. The last two phases are executed interactively and iteratively. In the Template Selection phase, evaluation methods are developed to recognize high-quality sequence-template alignments. In the Model Generation phase, restraints retrieved from the alignments and models (generated in the previous iteration) are used to build new models by Multi-dimensional Scaling (MDS)[2] techniques. In the Model Selection phase, both single-model quality assessment (QA) and consensus QA methods are developed to evaluate the quality of models.

MUFOLD searches PDB to get sequence-template alignments by search engines such as PSI-BLAST[3], HHSearch[4] and in-house threading tools. With these alignment hits, a QA method is developed to evaluate their quality and select top ones for further model generation. The main idea is to calculate the fitness between the target and the aligned substructure of the templates, and select high-quality ones according to the distribution of the fitness scores of all alignments. The fitness score includes sequence similarity, the matches between the templates' SS (secondary structure), SA (solvent accessibility) and the predicted SS, SA of the target sequence, the consensus of one alignment hit to all the other hits, etc.

The selected top alignments are then clustered into different groups by structure-similarity comparison. The alignments in each group share some highly similar substructures (i.e., conserved regions) and also keep diversities (i.e., non-conserved regions). These regions are detected by a graph-based QA method. Various distance restraints are retrieved through sampling the above conserved and non-conserved regions, and models are built for each set of distance restraint through applying MDS techniques. By MDS, MUFOLD can accommodate diverse spatial restraints retrieved from heterogeneous alignments.

Model-level QA is very important for structure prediction as a model can provide more detailed information in 3D than sequence-template alignment. In the Model Selection phase, both single-model QA and consensus QA methods are used to evaluate and select good models. At first, single-model QA methods such as OPUS-CA[5], DDFire[6] and Model Evaluator Score[7] are applied to filter out poor models. The remaining models are evaluated by consensus QA method. For example, the models are clustered and the k (e.g., k=5) representatives of each cluster, which have the biggest average similarity to the other members in the cluster are selected as top models.

In MUFOLD, the model generation and selection phases are executed interactively and

iteratively. In particular, the restraints between residues are filtered and iteratively refined by combining the original restraints derived from the alignments (*Dalignment*) and the measured distances from the generated models (*Dmodel*) as *Drefine* = λ**Dalignment* + (1 –λ)* *Dmodel*, $0 \leq \lambda \leq 1$, where value of λ is decided by the graph-based QA method. By performing this iterative generation, the quality of models often becomes better and better, while many deficiencies in the models are fixed over iterations.

We also use the strategy of MUFOLD-Server on CASP10 human prediction and refinement prediction. Different from the Server prediction which uses the PDB structures searched by search engines as templates, human prediction uses all CASP10 server prediction models as templates while refinement prediction uses the starting model and the PDB structures which are close to the starting model as templates. A preliminary assessment[8] on the CASP10 targets with release structures shows that MUFOLD-Server performs much better than it did in CASP9 and CASP8.

Availability

1. Zhang,J., Wang, Q., Barz, B., He, Z., Kosztin, I., Shang, Y., & Xu, D.(2009). MUFOLD: A New Solution for Protein 3D Structure Prediction. *Proteins: Struct Funct Bioinformatics.* **78**(5), 1137-52.
2. Borg, I. & Groenen, P. (1997). Modern multidimensional scaling—theory and applications. *New York: Springer-Verlag*.
3. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
4. Soding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics.* **21**, 951–960.
5. Wu, Y., Lu, M., Chen, M., Li, J. & Ma, J. (2007) OPUS-Ca: a knowledge-based potential function requiring only Ca positions. *Protein Sci.* **16**, 1449–1463.
6. 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.
7. Wang, Z., Tegge, A. & Cheng, J. (2009) Evaluating the absolute quality of a single protein model using structural features and support vector machines. *Proteins: Struct Funct Bioinformatics*. **75**, 638–647.
8. http://zhanglab.ccmb.med.umich.edu/casp10/

Refinement and Selection of Near-native Protein Structures

Jiong Zhang[1], J. Zhang[2,3], D. Xu[2,3], Y. Shang[2], I. Kosztin[1]

[1]*Department of Physics and Astronomy,* [2]*Department of Computer Science* [3]*Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211, USA*
KosztinI@missouri.edu

In recent years *in silico* protein structure prediction reached a level where a variety of servers can generate large pools of near-native structures. However, the identification and further refinement of the best structures from the pool of decoys continue to remain problematic. To address these issues, we have developed a *selective refinement* protocol (based on the Rosetta software package) and a Molecular Dynamics (MD) simulation based ranking method (MDR). The refinement of the selected structures is done by employing Rosetta's relax mode, subject to certain constraints. The selection of the final best models is done with MDR by testing their relative stability against gradual heating during all atom MD simulations. We have implemented the selective refinement protocol and the MDR method in our fully automated server Mufold-MD, which follows three sequential steps: 1) model generation, 2) selective refinement, and 3) MDR selection. We have tested Mufold-MD in the CASP10 competition.

For model generation, Mufold-MD uses 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. For hard targets models (~8,000) are generated using the Rosetta 3.3 software[1-4] (*ab-initio* method). To this end, secondary structure information from the amino acid sequence is obtained with PSIPRED[5] and fragment libraries are built from the NCBI database files. For further processing only the N lowest Rosetta energy structures are retained. For easy targets models (~2,000) are generated by using the Multi-dimensional Scaling (MDS) method[6], and subsequently ranked according to their OPUS_Ca[7] scores. Again, the top N structures are retained for further refinement (using Rosetta 3.3).

For selective refinement, based on their structure information, targets are divided into different categories and subjected to appropriate constraints. Targets that contain only α-helices or β-sheets are refined without constraints. The refinement of large and complicated targets is done by fixing their Ca atoms and leaving their side-chain atoms unconstrained. In moderately complex large targets one identifies stable substructures and keep fixed the corresponding Ca atoms during refinement. The rest of the targets are refined by applying standard deviation weighted constraints to the Ca atoms. For each model, a small number n (~10) of refined structures are generated. From the $n \times N$ refined structures only the top N models (with the lowest Rosetta energy) are retained.

Finally, from the N refined structures, the top 5 models are selected with the MDR method. For the MD simulations, first, the missing hydrogen atoms are added to the structures by using PSFGEN, which is part of the visual molecular dynamics (VMD) package[8]. Next, the 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 our energy

minimizations and MD simulations were performed with the parallel NAMD2.8 MD simulation program[12], by employing the CHARMM force field[9,10] (for β-sheet dominated targets) or Amber force field[11] (for α-helices dominated targets). 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 CASP10 competition. For CASP10, the decoys were generated on 47 dual-core Intel Xeon EM64T-2.8GHz CPUs, and *N*=94. Once the native structures for the CASP10 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 CASP10 meeting.

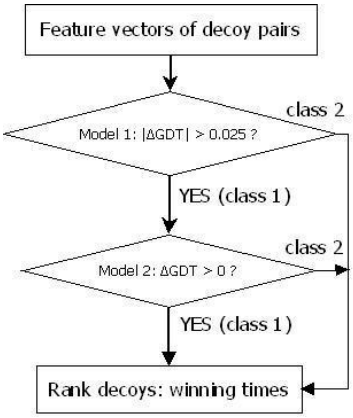
1. Bonneau, R., Strauss, C. E. M., Rohl, C. A., Chivian, D., Bradley, P., Malmström, L., Robertson, T. & Baker, D. (2002) *Journal of Molecular Biology* **322,** 65-78.
2. Bonneau, R., Tsai, J., Ruczinski, I., Chivian, D., Rohl, C., Strauss, C. E. M. & Baker, D. (2001) *Proteins: Structure, Function, 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. Zhang, J., Wang, Q., Barz, B., He, Z., Kosztin, I., Shang, Y. & Xu. D. (2009) Proteins: Structure, Function, and Bioinformatics **78**, 1137-1152.
7. Wu Y, Lu M, Chen M, Li J, Ma J (2007) *Protein Sci 2007* **16(7),** 1449-1463.
8. Humphrey, W., Dalke, A. & Schulten, K. (1996) *J. Mol. Graphics* **14,** 33-38.
9. MacKerell Jr, A. D., Bashford, D., Bellott, M. & others (1992) *FASEB J.* **6,** A143-A143.
10. MacKerell Jr, A. D., Bashford, D., Bellott, M. & others (1998) *J. Phys. Chem. B* **102,** 3586--3616.
11. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould & others (1995) Journal of the American Chemical Society **117**:5179-5197
12. 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.

Combining Consensus GDT with Single Scoring Functions for Selection of Near Native Structures

Zhiquan He, Alazmi Meshari , Jingfen Zhang and Dong Xu

Department of Computer Science and Christopher S. Bond Life Sciences Center,
University of Missouri, MO 65211, USA
xudong@missouri.edu

Selection of near native decoys is still a challenging problem for protein structure prediction. Consensus Global Distance Test (CGDT) has proved to work well when good decoys are in a majority cluster, which is specially the case in CASP. However, single scoring functions have their own merits as CGDT only considers the geometry information from the decoy set. To address this issue, we developed a method to combine single scoring functions and consensus GDT, and applied it to the QA session in CASP10, as shown in the flowchart. The basic idea is to compare any two decoys in terms of their structure quality first and then combine all the comparisons for QA of each decoy. First, the difference between feature vectors of a decoy-pair A and B were input to two independent neural network models to decide whether A or B is closer the native structure, in terms of GDT score. The first model was to judge whether two decoys are significantly different. If yes, the second model was used to decide which one of the two was better.



The feature vector for each decoy included

1. *CGDT*
2. Secondary structure match score
3. Solvent accessibility match score
4. Mean square error between predicted angles and actual decoy angles
5. Structural environment fitness score between sequence and decoy structure

After the pairwise comparison between all decoy pairs, the final score, named as PWCom, for each decoy was the number of winning times during the pair-wise comparison. If a decoy-pair falls in class 2 from model 1, they were close enough to be treated as identical.

The neural networks were trained and tested on CASP9. Similar method was also applied to the I-TASSER data set, which contained 56 targets. Table 1 showed the comparison of CGDT and PWCom in terms of top-1, top-5 selection performance and their Spearman correlation to the actual GDT score.

	CASP 9, 44 Targets			Yang Zhang's, 56 Targets		
	*top*1	$\overline{top5}$	Spearman	*top*1	$\overline{top5}$	Spearman
GDT	0.6412	0.6243	1.0000	0.6946	0.6767	1.000
CGDT	0.5861	0.5851	0.8408	0.6058	0.6039	0.5845
PWCom	**0.5958**	**0.5904**	**0.8499**	**0.6105**	**0.6056**	**0.6011**

Table 1: Performance in top-1, average top-5 selection and correlation

In conclusion, test results show that combination of consensus GDT and single scoring functions improves over the naïve consensus GDT method in selection performance and correlation. Further improvements can be achieved by choosing better single scoring functions and parameter

settings used in the method.

1. Cheng, J., Randall, A. Z., Sweredoski, M. J. & Baldi, P. (2005). SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Res* **33**, W72-6.
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**, 847-56.
3. Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577-637.

The MULTICOM Conformation Ensemble Approach to Protein Tertiary Structure Prediction

Xin Deng, Jilong Li, Badri Adhikari, and Jianlin Cheng*

Department of Computer Science, University of Missouri, Columbia, MO 65211, USA

*chengji@missouri.edu

Our four tertiary structure prediction servers (MULTICOM-NOVEL, MULTICOM-REFINE, MULTICOM-CLUSTER, MULTICOM-CONSTRUCT) and one human tertairy structure predictor (MULTICOM) participated in the CASP10 experiment. Despite of implementaiton differences and new developoments, they were based on the same *conformation ensemble* approach to protein tertairy structure prediction[1,2].

Methods

The basic *conformation ensemble* protocol[1,2] in our server predictors generated an ensemble of protein models for each target using multiple templates identified by dozens of sequence/profile comparison tools (e.g., BLAST, PSI-BLAST, HHSearch[3], SAM[4], HMMer) and fold recognition tools, alternative target-template alignments, and three complementary model generation tools. The ensemble of hundreds (e.g., 150 – 250) of models generally approximated the near native conformations of the target well if one or more homologous templates were identified for the target. For some hard targets for which no good template was found, tens of models selected from hundreads of models generated by a fragment assembly based approach (i.e. Rosetta[5]) were added into the ensemble in order to increase the diversity of the model pool as well as the frequency of near native conformation fragments.

The ensemble of models of a target were evaluated by several different methods, including the single-model absolute model quality assessment tool – ModelEvaluator[6], the fully pairwise model comparision tool – APOLLO[7], a protein energy calculation tool – SELECTpro[8], and the frequency of the templates (i.e., number of times that a template was chosen by different sequence/profile comparison tools) used to generated models if any. From the ensemble, MULTICOM-REFINE selected top five models ranked by APOLLO quality scores, which may subject to further multiple model combination and/or ab initio tail refinement based on a recursive protein modeling (RPM) protocol; MULTICOM-CLUSTER selected top five models ranked by the consensus ranking of ModelEvaluator scores and SELECTpro energies; MULTICOM-NOVEL considered both APOLLO scores and template frequence in model selection; and MULTICOM-CONSTRUCT chose top five models ranked by the sum of APOLLO scores and ModelEvaluator scores. For multi-domain targets, APOLLO may be run on individual domains to rank and select top domains to combine into full models. For human prediction, our method MULTICOM used all the CASP server predictions as the model ensemble because the CASP pool of models generated by 60+ server predictors formed a better approximation of the near native conformations of a target than our own server models. All the models in the human ensemble were evaluated by the four measures described above, our new weighted pairwise model comparision method (see our MULTICOM-CONSTRUCT QA abstract) and our new domain-based single-model and clustering-based model evaluation methods. The top models selected by the consensus of the complementary ranking metrics and/or human inspection were refined by the model / domain combination protocol[1,2] and then

submitted to CASP by MULTICOM.

In comparison with our methods tested in CASP9, the major new developments fully or partially benchmarked in CASP10 include: 1) a new fold recognition method based on information propagation on the pairwise sequence / structure similarity network of template protiens; 2) a newly developed multiple sequence aligment tool (MSACompro[9]) based on sequence profile, predicted secondary structure, and solvent accessibility that was used to align a target with multiple templates to generate more alternative alignments for model generation; 3) a new in-house template-based model generation tool that constructed the core structure for a target from templates and filled in the unaligned region (e.g. loop) with fragments of variable length extracted from a database of representive proteins according to both sequence similairty and structural fitness with the core structure; and 4) new domain-based model evaluation methods, weighted pairwise comparison-based model evaluation methods, and an alignment-based model evaluation method, which were tested with at least one of our predictors on some or all targets informing model selection during the CASP10 experiment.

Results

We evaluated preliminarily our four server predictors on the whole chains of 34 CASP10 targets whose experimental structures were released to date. **Table 1** reports the average GDT-TS scores and TM-scores of top 1 or 5 models predicted by these predictors.

Table 1. The average GDT-TS scores and TM-scores of top one and best of five models on 34 targets.

Predictors	Top One			Best of Five	
	GDT-TS	TM-score		GDT-TS	TM-score
MULTICOM-NOVEL	0.5418	0.6239		0.5530	0.6340
MULTICOM-REFINE	0.5417	0.6236		0.5526	0.6357
MULTICOM-CLUSTER	0.5353	0.6175		0.5526	0.6352
MULTICOM-CONSTRUCT	0.5345	0.6183		0.5477	0.6326

1. Wang Z, et al. (2010). MULTICOM: A multi-level combination approach to protein structure prediction and its assessment in CASP8. Bioinformatics. 26(7):882-888.
2. Cheng J, Li J, Wang Z, Eickholt J, Deng X. (2012). The MULTICOM Toolbox for Protein Structure Prediction. BMC Bioinformatics, 13:65.
3. Söding J. (2005). Protein homology detection by HMM–HMM comparison. Bioinformatics, 21(7):951.
4. Karplus K, et al. (1997). Predicting protein structure using hidden Markov models K. Karplus, K. Proteins, S1:134--139.
5. Leaver-Fay A, et al. (2011). ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol, 487: 545-74.
6. Wang Z, Tegge AN, Cheng J. (2009). Evaluating the absolute quality of a single protein model using support vector machines and structural features. Proteins, 75(3):638-647.
7. Wang Z, Eickholt J, and Cheng J. (2011). APOLLO: a quality assessment service for single and multiple protein models. Bioinformatics, 27(12), 1715-1716.
8. Randall A, Baldi P. (2008). SELECTpro: effective protein model selection using a structure-based energy function resistant to BLUNDERS. BMC Structural Biology, 8:52.
9. Deng X, Cheng J. (2011). MSACompro: protein multiple sequence alignment using predicted secondary structure, solvent accessibility, and residue-residue contacts. BMC Bioinformatics, 12:472.

Protein Residue-Residue Contact Prediction by the MULTICOM Predictors

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

Methods

MULTICOM-CONSTRUCT is a new sequence based contact predictor built using boosted ensembles of *deep networks* (DNs). For medium and long-range contacts, the ensemble consists of 490 DNs. The overall architecture of each DN is X-500-500-350-1 with the size of the input layer X depending on the width of two fixed input windows centered on the residue pair to be classified. Each DN is trained layer by layer using contrastive divergence[1] and then fine tuned using standard back propagation. Boosting was accomplished by sampling the training data for each DN from a large pool of training examples. Initially, the probability of including a training example was uniformly distributed and then adjusted after each round of boosting. More specifically, the training examples which were misclassified by the previously trained DN had its probability of selection increased while correctly classified training examples had their probability decreased. For short-range contacts, the ensemble consisted of 30 DNs with an overall architecture of 400-500-500-250-21. The input for each DN in this setting came primarily from one window 12 residues in length and the target was all short-range contact pairings contained in the input window. The features used as input included predicted secondary structure and solvent accessibility, values from a position specific scoring matrix, and a number of statistical pair wise potentials.

MULTICOM-CLUSTER is a sequence based, *ab-initio* predictor based on our residue-residue prediction tool SVMcon[2]. This approach used a support vector machine (SVM) to classify residue-residue pairings. The input to the SVM consisted of features such as predicted secondary structure, predicted solvent accessibility and a sequence profile for residues contained in two 9-residue long windows centered on the residue pair in question. The SVM was trained on a large dataset and classified each residue pair as "in-contact" or "non-contact". Those pairs classified as "in-contact" were submitted as predictions.

MULTICOM-NOVEL and MULTICOM-REFINE are sequence-based, *ab-initio* methods based on our recursive neural network predictor, NNcon[3]. 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 (i.e., MULTICOM-NOVEL used only general residue-residue contact predictions, whereas MULTICOM-REFINE combined specific beta-residue contact predictions with general residue contact predictions). Features used for each residue include a sequence profile, predicted secondary structure and solvent accessibility.

Finally, our human predictor MULTICOM used an automated, model conformation ensemble approach to make residue-residue contact predictions[4]. The method works by consolidating residue-residue contacts from a number of structural models generated for a target. In this case, we used the full set of models submitted by those participants in the server category. MULTICOM used a consensus voting approach which extracted contacts from all the tertiary structure models 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 contacts. A principle advantage of this approach is the ability to consolidate contact information across models regardless of conformation.

Results

As an initial assessment of our techniques on the CASP10 dataset, we evaluated the methods on 36 valid CASP targets. Tables 1 and 2 show the accuracy and coverage of the top L and L/5 long and medium range contacts where L is the length of the protein. Note that this evaluation is done on a per target basis and irrespective of the domain architecture.

Table 1. Preliminary results of the MULTICOM series on long range contact predictions on 36 valid CASP10 targets.

	Top L/5 long range		Top L long range	
Method	Acc.	Cov	Acc.	Cov.
MULTICOM-CONSTRUCT	0.231	0.038	0.141	0.115
MULTICOM-CLUSTER	0.142	0.023	0.093	0.075
MULTICOM-REFINE	0.151	0.024	0.084	0.069
MULTICOM-NOVEL	0.144	0.024	0.083	0.068

Table 2. Preliminary results of the MULTICOM series on medium range contact predictions on 36 valid CASP 10 targets.

	Top L/5 medium range		Top L medium range	
Method	Acc.	Cov.	Acc.	Cov.
MULTICOM-CONSTRUCT	0.398	0.144	0.242	0.440
MULTICOM-CLUSTER	0.311	0.112	0.194	0.336
MULTICOM-REFINE	0.362	0.131	0.224	0.387
MULTICOM-NOVEL	0.358	0.129	0.222	0.382

Availability

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

1. Hinton, G.E. (2002). Training Products of Experts by Minimizing Contrastive Divergence. *Neural Computation* **14**, 30p.

2. Cheng, J. & Baldi, P. (2007). Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinformatics*. **8**,113
3. 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
4. Eickholt, J., Wang, Z., & Cheng, J. (2011). A Conformation Ensemble Approach to Residue-Residue Contact. *BMC Structural Biology*, **11**,38.

Contact Assisted Protein Structure Prediction by MULTICOM-CONSTRUCT

Badri Adhikari, Xin Deng, Jilong Li, Debswapna Bhattacharya, and Jianlin Cheng

Department of Computer Science, University of Missouri, Columbia, MO 65211 USA
chengji@missouri.edu

Our server MULTICOM-CONSTRUCT participated in the CASP10 contact assisted structure prediction. The experimental server was developed and continuously updated during the CASP10 prediction season. Although the server accidently missed three targets, it still generated some valuable, first-hand insights about how to use extra contact information to improve model construction and selection.

Method

Our approach for contact assisted protein structure prediction comprised of six major steps: 1) collect CASP10 server models; 2) use APOLLO[1] to assess these models and score them; 3) score the models based on how well they satisfy the contacts or no-contacts; 4) rank the models by integrating the scores obtained in steps 2 and 3 (i.e. Apollo's GDT-TS score, Apollo's MaxSub score, Apollo's TM-score, percent of exact contacts satisfied, percent of no-contacts satisfied); 5) select top 5 models and refine them using 3Drefine[2]; and 6) perform contact assisted structure prediction using Modeller[3] with contacts as distance restraints. Step 6 was added in the middle of the CASP10 experiment, and thus was only applied to some targets. Specifically, for each target, a pool of tertiary structure models was downloaded from the CASP10 web site. The pairwise model comparison based tool APOLLO was used to evaluate each model, resulting in three quality scores in terms of GDT-TS score, MaxSub score, and TM-score. The models were also scored based on what percent of given contacts they satisfied (i.e. number of given contacts present in the model divided by total number of given contacts). In cases when no-contacts rather than contacts were provided, the models were negatively scored based on what percent of no-contacts they satisfied. To calculate the total score for each model, the following formula was used: *Total score = APOLLO's GDT-TS score + APOLLO's MaxSub score + APOLLO's TM-score + percent of contacts satisfied – 0.1 * percent of no-contacts satisfied*.

The top models ranked by the total scores were refined by 3Drefine. To generate the final models, refined models were provided as templates for Modeller along with contacts as distance restraint. The contacts information was coded as distance restrains between Cα-Cα atoms (or Cβ atoms in case of GLY residue) using a harmonic potential function with 8.0 angstrom mean distance and 0.1 standard deviation.

Results

The final predicted models of contact assisted prediction target were evaluated based on how well they satisfied the contacts. Table 1 compares the percent of contacts satisfied by the top 1 model selected from downloaded models using total score against that by the top 1 model generated by our contact assisted prediction pipeline. The results showed that remodeling the top models using given contacts with Modeller improved or did not change the percent of contact satisfaction in all but one case.

For eight targets whose experimental structures were known by the time of writing this abstract, TM-score[4] was used to evaluate the predicted models. Table 2 shows that, in six out of

eight cases, the TM-scores of the final predicted models are higher than the initial models.

Target	Initial % contacts match	Final % contacts match	Improvement
Tc649	0.125	0.125	0
Tc676	0.118	0.118	0
Tc653	-	-	-
Tc658	0.188	0.188	0
Tc678	0.333	0.417	0.084
Tc673	0.200	0.200	0
Tc666	0.214	0.357	0.143
Tc691	0.400	0.400	0
Tc684	0.125	0.125	0
Tc680	0.444	0.222	-0.222
Tc734	0.200	0.200	0
Tc705	0.118	0.206	0.088
Tc717	0.200	0.267	0.067
Tc719	0.231	0.231	0
Tc735	0.229	0.343	0.114

Table 1. Percentage of contacts satisfied by the top 1 model for each target before and after contact-assisted modeling. Initial % contacts match is the percentage of contacts satisfied by the top 1 model selected from downloaded models using total scores. Final % contacts match is the percentage of contacts satisfied by the top 1 model predicted by the contacts-assisted pipeline. Improvement is the difference between the final % contacts match and the initial % contacts match. For target Tc653 no-contacts were provided instead of contacts. It is worth noting that, even though all the final models evaluated here were generated by the complete prediction pipeline with all the six steps during the CASP prediction season, the models of targets Tc734, Tc705, and Tc717 were not submitted to CASP by mistake, the models of targets Tc649, Tc676, Tc653, Tc658, Tc678, Tc673, and Tc666 actually submitted to CASP did not go through Step 6 of the pipeline as the step was added after the targets expired, and the initial models of target Tc649 used to generate models actually submitted to CASP were selected by the program with a bug in calculating the percent of contact match.

Target	Initial TM-Score	Final TM-Score	Improvement
Tc649	0.360	0.370	0.0099
Tc676	0.322	0.334	0.0114
Tc658	0.758	0.767	0.0081
Tc678	0.447	0.455	0.0083
Tc673	0.292	0.284	-0.0082
Tc680	0.728	0.607	-0.1211
Tc705	0.348	0.367	0.0192
Tc735	0.335	0.432	0.0971

Table 2. Evaluation of the top 1 predictions of the targets whose experimental structures were released. Initial TM-Score is the TM-Score of top 1 initial model selected by total score. Final TM-Score is the TM-Score of the top 1 model predicted by the six-step prediction pipeline. Improvement is the difference between the final TM-Score and the intial TM-Score.

Availability

The server is available at: http://protein.rnet.missouri.edu/contact_assisted/

1. Wang Z, Eickholt J, and Cheng J (2011). APOLLO: A Quality Assessment Service for Single and Multiple Protein Models, Bioinformatics, 27(12), 1715-1716.
2. Bhattacharya, D. and Cheng J (2012). 3Drefine: Consistent protein structure refinement by optimizing hydrogen bonding network and atomic-level energy minimization. Proteins: Structure, Function, and Bioinformatics.
3. N. Eswar, M. A. Marti-Renom, B. Webb, M. S. Madhusudhan, D. Eramian, M. Shen, U. Pieper , A. Sali (2006). Comparative Protein Structure Modeling With MODELLER. Current Protocols in Bioinformatics, John Wiley & Sons, Inc., Supplement 15, 5.6.1-5.6.30.
4. Zhang Y, Skolnick J (2005). TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic acids research, 33(7):2302-2309.

Protein Structure Refinement by Two-Step Atomic-level Energy Minimization

Debswapna Bhattacharya and Jianlin Cheng*

Department of Computer Science, University of Missouri, Columbia, MO 65211, USA

*chengji@missouri.edu

We participated in CASP10 refinement experiment with our automated refinement server MULTICOM-CONSTRUCT with the goal to improve model qualities consistently over the initial models.

Methods

MULTICOM-CONSTRUCT is our recently developed refinement protocol [1] which refines protein structures in two steps: (1) Optimizing Hydrogen Bonds network and (2) atomic-level energy minimization using a combination of physics and knowledge based force fields. In the first step, a search is performed for the polar hydrogen atoms to find out the most favorable position of hydrogen atoms satisfying hydrogen bonds with the closest neighboring atoms and considering the protonation state of each amino acid in order to optimize the hydrogen bonding network in the starting model. We call this "extended atomic model". The total potential energy of the extended atomic model is then computed using a combination of physics based and knowledge based force fields. Finally, a limited memory Broyden-Fletcher-Goldfarb-Shannon (L-BFGS) algorithm [2] is employed to minimize the total potential energy using MESHI molecular modeling package [3]. The energy minimized model is the refined structure.

Results

The preliminary assessment of MULTICOM-CONSTRUCT is based on 10 refinement targets for which the native structures have been released at the time of writing this abstract. We employ a two-fold evaluation technique to test the ability of our method to perform simultaneous improvement in both global and local model qualities. TM-score program [4] is used to compare the RMSD, GDT-HA and TM-score before and after refinement against the native structures to access the improvement in the global positioning of the backbone C_α atoms. In order to evaluate the enhancement of local structural qualities, we use MolProbity score [5] which takes into account the rotamer outliers, torsion-angle outliers, and steric clashes that have values outside the region of experimentally derived standard protein structures.

Table I reports the cumulative change in GDT-HA, TM-score, RMSD score and MolProbity scores for the best of five and the first submissions for all 10 refinement targets. Overall, there are 8, 8, 9 and 6 cases when MULTICOM-CONSTRUCT improves the GDT-HA, TM-score, RMSD and MolProbity scores respectively based on best of five submissions. For the top submitted models, the instances for successful refinement are 7, 7, 8 and 5 with respect to GDT-HA, TM-score, RMSD and MolProbity scores respectively. The ability of our method to successfully rank the best submitted model is also encouraging with the best overall refined model being submitted as model 1 for 6 out of 10 targets (60%). However, the refinement is often modest with improvement in GDT-HA score in the range of 0.01 to 0.03. The most promising aspect of this protocol is, therefore, consistency. More than 70% of the times, our

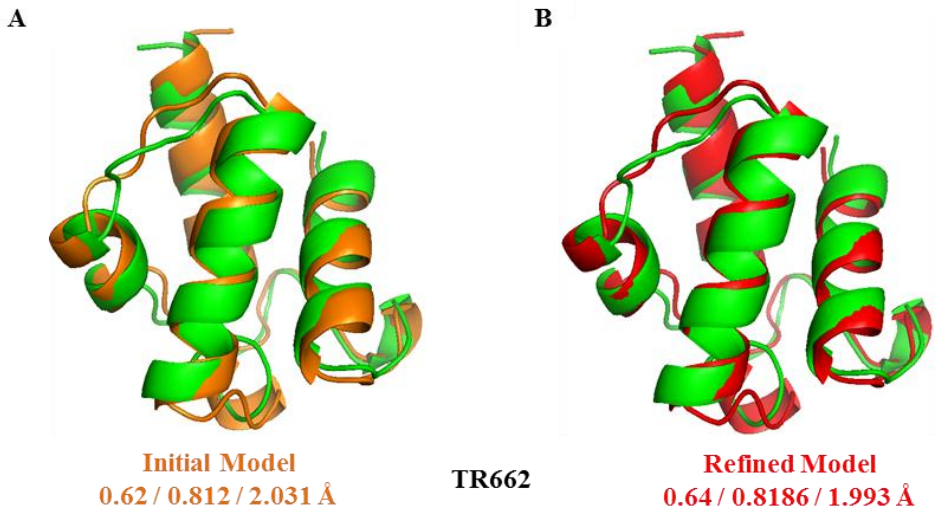


Figure 1. Example of refinement for target TR662.
(A) Structural superposition of initial model (orange) on native structure (green). The values under the model indicate GDT-HA, TM-score and RMSD score respectively before refinement.
(B) Structural superposition of refined model using MULTICOM-CONSTRUCT (red) on native structure (green). The values under the model indicate GDT-HA, TM-score and RMSD score respectively after refinement.
Figures were prepared in PyMOL (The PyMOL Molecular Graphics System, Version 1.4.1, Schrödinger, LLC.).

ent category*.

Submission	No. of Targets[a]	Σ (Δ GDT-HA)[b]	Σ (Δ TM-score) [c]	Σ (Δ RMSD) [d]	Σ (Δ MolProbity) [e]
Best of Five[f]	10	0.0581	0.0147	0.137	0.019
Top One[g]	10	0.0321	0.0072	0.069	1.101

* The numbers represents the cumulative change in score for each metric before and after refinement. A positive number indicates that the quality of the metric is improved by refinement and a negative number indicates degradation in quality corresponding to that metric after refinement.
[a] Number of targets submitted in CASP10 refinement experiment.
[b] Cumulative change in GDT-HA score.
[c] Cumulative change in TM-score.
[d] Cumulative change in RMSD score.
[e] Cumulative change in MolProbity score.
[f] First submitted model by MULTICOM-CONSTRUCT.
[g] Best of five submission by MULTICOM-CONSTRUCT.

method can improve the qualities of the starting structures.

A typical example of refinement is shown in **Figure 1** for the target TR662 based on the best submitted refined model by MULTICOM-CONSTRUCT. The initial model is quite accurate with RMSD of 2.031 Å. After refinement, the RMSD is improved to 1.993 Å along with modest but consistent improvement in GDT-HA and TM-score.

Availability

MULTICOM-CONSTRUCT web server (i.e. 3Drefine server) is freely available at http://sysbio.rnet.missouri.edu/3Drefine/.

1. Bhattacharya, D. and J. Cheng, *3Drefine: Consistent protein structure refinement by optimizing hydrogen bonding network and atomic-level energy minimization.* Proteins: Structure, Function, and Bioinformatics, 2012.
2. Liu, D.C. and J. Nocedal, *On the limited memory BFGS method for large scale optimization.* Mathematical programming, 1989. **45**(1): p. 503-528.
3. Kalisman, N., et al., *MESHI: a new library of Java classes for molecular modeling.* Bioinformatics, 2005. **21**(20): p. 3931-3932.
4. Zhang, Y. and J. Skolnick, *Scoring function for automated assessment of protein structure template quality.* Proteins: Structure, Function, and Bioinformatics, 2004. **57**(4): p. 702-710.
5. Chen, V.B., et al., *MolProbity: all-atom structure validation for macromolecular crystallography.* Acta Crystallographica Section D: Biological Crystallography, 2009. **66**(1): p. 12-21.

Protein Model Quality Prediction by MULTICOM Server Predictors

Renzhi Cao[1], Zheng Wang[1], Jilong Li[2], Charles Shang[4], 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. [4] – Rock Bridge High School, Columbia, MO 65203, USA
chengji@missouri.edu

Our group developed and tested four model quality assessment (QA) servers: MULTICOM-REFINE, MULTICOM-CLUSTER, MULTICOM-NOVEL, MULTICOM-CONSTRUCT. They predicted both global quality scores and local quality scores for stage1 and stage 2 models of CASP10 targets.

Methods

MULTICOM-REFINE uses a pair-wise model comparison approach (APOLLO)[1] to generate the global quality score. The 19 top models based on the global quality scores and the top 1 model selected by SPICKER[2] formed a top model set for local quality prediction. After superimposing predicted model with each model in top model set, the local quality score is calculated as the average absolute difference between each residue in the model and the residue in the model from top model set.

MULTICOM-CLUSTER is a new, single-model, support vector machine (SVM)-based method. The input features to the SVM includes amino acids encoded by a 20-digit vector of 0 and 1, the difference between secondary structure and solvent accessibility predicted by SCRATCH[3] from the protein sequence and that of a model parsed by DSSP, and predicted contact probabilities. The SVM was trained to predict the local quality score of each residue. The predicted local quality score was used to generate the global quality score of the model according to the formula:

$$Global\ quality\ score = \frac{1}{t}\sum_{i=1}^{t}\left(\frac{1}{1+(\frac{L_i}{p})^2}\right).$$

In the formula, t is the total number of residues, L_i is the local quality score of residue i, and p is a parameter whose value is set to 5. Residues that didn't have predicted local quality scores were skipped in averaging.

MULTICOM-NOVEL is the same as MULTICOM-CLUSTER except that amino acid sequence features were replaced with the sequence profile features. The multiple sequence alignment of a target used to generate profiles was generated by PSI-BLAST.

MULTICOM-CONSTRUCT is a weighted pairwise model evaluation approach to predict global quality. It uses ModelEvaluator [4] – an ab initio single-model global quality prediction method – to predict a score for each model and TM-SCORE to get the GDT-TS score for each pair of models. The predicted global quality score of a model i is the weighted average GDT-TS score between the model and other models, calculated according to the formula: $S_i = \sum_{j=1}^{N}\left(X_{i,j} * \frac{W_j}{\sum_{j=1}^{N} W_j}\right)$. In this formula, S_i is the predicted global quality score for model i, N is the total number of models, $X_{i,j}$ is the GDT-TS score between model i and model j, W_j is the

score for model *j* predicted by ModelEvaluator. In case that no score was predicted for a model by ModelEvaluator, the weight of the model was set to the average of all the scores predicted by ModelEvaluator. The local quality prediction of MULTICOM-CONSTRUCT is the same as MULTICOM-NOVEL except that additional SOV (segment overlap measure of secondary structure) score features were used by the SVM to generate the local quality score.

Results

We preliminarily evaluated the performance of our four servers on 33 CASP10 target structures released by the time of writing the abstract. We evaluated the global quality scores predicted by our four servers against the real quality scores according to two metrics: average per target correlation and average per target loss. The loss for each target was the difference between the GDT-TS score of the overall best model and the top model ranked by the global quality scores. For the local quality score, the correlation score for each model was calculated as the correlation of the predicted local quality scores for the residues in the model and the real local quality scores. The average correlation score for all models associated with a target was used as the correlation of the target. The performance of the four servers was reported in **Table 1**.

Table 1. The average per-target correlation score, average loss, average local quality correlation score for stage1 and stage2 models.

Server Predictors	Ave. Corr. Stage1 (Global)	Ave. loss Stage1 (Global)	Ave. Corr. Stage1 (Local)	Ave. Corr. Stage2 (Global)	Ave. loss Stage2 (Global)	Ave. Corr. Stage2 (Local)
MULTICOM-REFINE	0.6707	0.0539	0.6223	0.5312	0.0534	0.6621
MULTICOM-CLUSTER	0.5588	0.0798	0.2774	0.3756	0.0680	0.3341
MULTICOM-NOVEL	0.5463	0.0852	0.2976	0.3699	0.0693	0.3521
MULTICOM-CONSTRUCT	0.7448	0.0451	0.2970	0.5492	0.0503	0.3328

1 Larsson, P., Skwark, M. J., Wallner, B. & Elofsson, A. Assessment of global and local model quality in CASP8 using Pcons and ProQ. *Proteins: Structure, Function, and Bioinformatics* **77**, 167-172 (2009).

2 Zhang, Y. & Skolnick, J. SPICKER: A clustering approach to identify near-native protein folds. *Journal of computational chemistry* **25**, 865-871 (2004).

3 Cheng, J., Randall, A., Sweredoski, M. & Baldi, P. SCRATCH: a protein structure and structural feature prediction server. *NAR* **33**, W72-W76 (2005).

4 Wang, Z., Tegge, A. N. & Cheng, J. Evaluating the absolute quality of a single protein model using structural features and support vector machines. *Proteins: Structure, Function, and Bioinformatics* **75**, 638-647 (2009).

A Conformation Ensemble Approach to Protein Structure Refinement

Debswapna Bhattacharya[1] and Jianlin Cheng [1,2,3,*]
[1]*Department of Computer Science, University of Missouri, Columbia, MO 65211, USA*

Improving the overall fold of the initial models is one of the major challenges in protein structure refinement field. In CASP10, we tackled this problem by applying a recently developed conformation ensemble approach.

Methods

For each refinement target, we collected all the submitted server tertiary structure predictions for the corresponding targets. The complete archive of submitted models by all servers had been used as the ensemble. In the first step, the problematic regions (PRs) in the starting models were predicted using a newly developed consensus method. A novel hybrid model generation approach was then employed using template fed to Modeller [1] by combining the initial model and the PR replaced by structures from the ensemble. In the next step, the hybrid models were ranked by complementary single-model quality evaluation schemes followed by a weighted rank aggregation method to derive the "optimal ranking" through Cross-Entropy Monte Carlo algorithm (CE) [2] using the Kendall's Tau distance for partially ordered lists [3]. A maximum of top three models were chosen from the hybrid models ranked above the starting structure in the "optimal ranking" as templates to feed into automodel class of Modeller to derive the "improved model" for the PR using multiple template alignment.

We adopted an iterative refinement strategy to gradually improve the initial model with each PR getting improved in a single iteration. The PRs were sorted based on their length and longer PRs getting higher priority than shorter PRs. After each iteration, the "improved model" corresponding to a PR becomes the starting model for the next round of iteration aiming to improve the next PR. This process continues until all the PRs in the initial model are consumed. Finally, the local structural errors and general physicality of the final "improved model" were enhanced by our previously developed refinement method [4] in order to produce the refined structure.

Results

We perform preliminary assessment of our method on 10 refinement targets for which the native structures have been released at the time of writing this abstract. The refinement is evaluated from two perspectives: (1) similarity to the native structures and (2) physical reasonableness of the models. In order to judge how the overall fold in starting structures were improved by refinement, we use TM-score program [5] to compare the RMSD, GDT-HA and TM-score before and after refinement against the native structures. The enhancement of physical reasonableness and correction of the local errors are evaluated using MolProbity program [6].

Table I summarizes the cumulative GDT-HA, cumulative TM-score, average RMSD and average MolProbity score of the best of five submissions for all 10 refinement targets. The NULL group represents the initial models issued for refinement. Overall, a 21.6% improvement in the average RMSD has been observed with GDT-HA score getting improved for 7 out of 10 targets (70%). Promisingly, our method can consistently rank the best submitted model at the top.

For 6 out of 10 targets (60%), the best overall refined model was submitted as model 1.

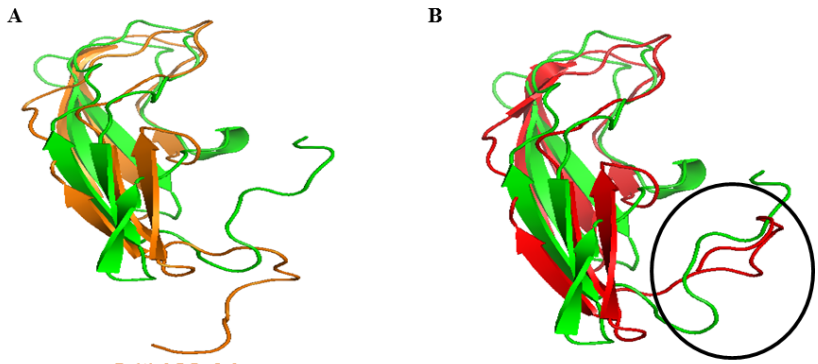


Figure 1. Example of refinement for target TR671.
(A) Structural superposition of initial model (orange) on native structure (green). The values under the model indicate GDT-HA, TM-score, RMSD and MolProbity score respectively before refinement.
(B) Structural superposition of refined model using MULTICOM-NOVEL (red) on native structure (green). The values under the model indicate GDT-HA, TM-score, RMSD and MolProbity score respectively after refinement. The black circle highlights the region with prominent structural improvement.
Figures were prepared in PyMOL (The PyMOL Molecular Graphics System, Version 1.4.1, Schrödinger, LLC.).

A representative example of refinement is shown in **Figure 1** for the target TR671 based on the first submitted refined model by MULTICOM-NOVEL. The initial model has an RMSD of 7.716 Å with a large deviation in the N-terminal region compared to the native structure. After refinement, the RMSD is improved to 5.008 Å with 2.3%, 2.5% and 14.7% improvement in GDT-HA, TM-score and MolProbity score respectively. The improvement in the N-terminal region is obvious by visual inspection.

Table I. Preliminary results for MULTICOM-NOVEL in CASP10 refinement category.

Group Name	No. of Targets[a]	GDT-HA[b]	TM-score[c]	RMSD[d]	MolProbity[e]
MULTICOM-NOVEL	10	5.552	7.860	3.01	2.5
Null[f]	10	5.530	7.769	3.66	2.6

[a] Number of CASP10 targets in the Refinement Experiment. [b] Cumulative GDT-HA score for best of five submissions. [c] Cumulative TM-score for best of five submissions. [d] Average RMSD for best of five submissions with respect to the native structure in Å. [e] Average MolProbity score for best of five submissions. [f] The initial models for the CASP10 refinement experiment.

Availability

The web server is freely available at http://sysbio.rnet.missouri.edu/REFINEpro/.

1. Fiser, A. and A. Šali, *Modeller: generation and refinement of homology-based protein structure models.* Methods in enzymology, 2003. **374**: p. 461-491.
2. Rubinstein, R., *The cross-entropy method for combinatorial and continuous optimization.* Methodology and computing in applied probability, 1999. **1**(2): p. 127-190.
3. Adler, L.M.K., *A modification of Kendall's tau for the case of arbitrary ties in both rankings.* Journal of the American Statistical Association, 1957. **52**(277): p. 33-35.
4. Bhattacharya, D. and J. Cheng, *3Drefine: Consistent protein structure refinement by*

optimizing hydrogen bonding network and atomic-level energy minimization. Proteins: Structure, Function, and Bioinformatics, 2012.
5. Zhang, Y. and J. Skolnick, *Scoring function for automated assessment of protein structure template quality.* Proteins: Structure, Function, and Bioinformatics, 2004. **57**(4): p. 702-710.
6. Chen, V.B., et al., *MolProbity: all-atom structure validation for macromolecular crystallography.* Acta Crystallographica Section D: Biological Crystallography, 2009. **66**(1): p. 12-21.

Protein Disorder Prediction by the MULTICOM Predictors

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

Three protein residue disorder predictors participated in CASP10 from the MULTICOM group. All three were fast, sequence based methods with two of the predictors using machine learning techniques and the other being a meta approach.

Methods

MULTICOM-NOVEL is a new approach to protein residue disorder prediction using *deep networks* (DNs) and boosted ensembles. To construct the predictor, we trained a number of DNs by sampling from a pool of training data. Initially, every example in the training pool had an equal probability of being included in the training sample. After each round of boosting, the probability of selecting properly classified examples was decreased while the probability of selecting a misclassified example was increased. A number of ensembles were trained using various DN architectures and the final prediction for a residue was a performance weighted sum of all of the DNs in the ensemble. Generally speaking the architecture used for the DN was X-750-750-350-Y where the size of the input window X varied from 644 to 964 for input windows of 20, 25 or 30 residues and the target window size of 3, 5 or 7. Note that inputs used as features included predicted secondary structure and solvent accessibility, values from a position specific scoring matrix, and a few statistical characterizations of the residues. Each DN was trained layer by layer using contrastive divergence[1] and the final weights fine tuned using a standard back propagation algorithm.

MULTICOM-REFINE made disorder predictions using 1-dimensional recursive neural network (1D-RNN) with input stemming from a sequence profile and predicted secondary structure and solvent accessibility[2]. The predicted disorder probabilities were rescaled such that the ratio of residues with a probability of disorder greater than or equal to 0.5 was similar to the ratio of disordered residues in the training set[3].

MULTICOM-CONSTRUCT is a fast, sequence based meta method which combines the predictions of both MULTICOM-REFINE and MULTICOM-NOVEL. Our initial evaluation and comparison of both MULTICOM-NOVEL and MULTICOM-REFINE indicated that the disorder residues identified by both methods are complementary at times and this lead to the construction of a simple meta approach between the two.

Results

As an initial assessment of our methods, we evaluated the disorder predictions of MULTICOM-CONSTRUCT, MULTICOM-REFINE and MULTICOM-NOVEL on 30 valid CASP10 targets which were solved using X-ray crystallography and available by the time of writing this abstract. Any residue which did not have coordinates specified in the PDB file was considered to be

disordered. In all there were a total of 9102 residues in our evaluation set with 978 of them being disordered. The results are summarized in Table 1.

Table1. Results of our MULITCOM series of disorder predictions on 30 valid CASP10 targets. AUC is the area under the ROC curve and ACC is the balanced accuracy.

Method	AUC	ACC
MULTICOM-NOVEL	0.814	0.771
MULTICOM-CONSTRUCT	0.814	0.751
MULTICOM-REFINE	0.771	0.728

Availability

MULTICOM-NOVEL (i.e., DNdisorder) is available as a web service at http://iris.rnet.missouri.edu/dndisorder/. MULTICOM-REFINE (i.e., PreDisorder) is available at http://casp.rnet.missouri.edu/predisorder.html as a web service and as downloadable software.

1. Hinton, G.E. (2002). Training Products of Experts by Minimizing Contrastive Divergence. *Neural Computation **14***, 30p.
2. Deng, X., Eickholt, J., & Cheng, J. (2009) PreDisorder: AbInitio Sequence-based Prediction of Protein Disordered Regions. *BMC Bioinformatics*. **10**,436.
3. Hecker, J., Yang, J. & Cheng, J. (2008). Protein Disorder Prediction at Multiple Levels of Sensitivity and Specificity. *BMC Genomics*. **9**, S1.

OnD-CRF2: Disorder prediction in proteins using Conditional Random Fields

L. Wang[1, 2] and U.H. Sauer[1, 2]

[1] - Department of Chemistry, [2] - Computational Life Science Cluster, CLiC, Umeå University, Sweden

uwe.sauer@chem.umu.se

An increasing number of proteins transfer key biological functions through intrinsically unstructured sequence intervals[1-2]. 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[3].

OnD-CRF2 is a newer version of OnD-CRF[4] for accurate prediction of **O**rdered a**n**d **D**isordered amino acid regions in proteins by using **C**onditional **R**andom **F**ields (CRF). The CRF models depend 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.

Methods

OnD-CRF2 was trained on a new training dataset including a set of around 5367 non-redundant sequences with high resolution X-ray structures. Disorder was identified with those residues that appear in the sequence records but with coordinates missing from the electron density map.
Performance is optimized with respect to the Area Under the ROC Curve (AUC) and the average of sensitivity and specificity (ACC), which are the measures of the overall predictor quality.
The OnD-CRF2 method makes use of the open source package CRF++ (http://crfpp.sourceforge.net/) to implement Conditional Random Fields (CRF). The template file used for training the OnD-CRF2 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[5]. We use cross-validation to optimize the parameters for CRF++ in order to generate the OnD-CRF2 model that achieves the maximal AUC and ACC values.

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 0.834 are: 1.0 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 $P < 0.10$ for ordered and $P \geq 0.10$ for disordered amino acids(*). Using this cut-off the OnD-CRF2 model achieves an ACC of 0.809 which are about 2 points per cent higher than the old version of OnD-CRF.

Availability

OnD-CRF2 server: http://babel.ucmp.umu.se/ond-crf2/

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. Wang L, Sauer UH. (2008)OnD-CRF: predicting order and disorder in proteins using conditional random fields. *Bioinformatics,* **24(11),** 1401-1402
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.

A Novel Procedure for Constructing Multiple Alignments for Protein Structure Prediction

K. Tomii
Computational Biology Research Center,
National Institute of Advanced Industrial Science and Technology (AIST)
k-tomii@aist.go.jp

We have developed template-based protein structure modeling protocols based on our profile-profile alignment method[1-2]. However, constructing accurate alignment(s) between a target and a template is still one of the central issues in protein structure prediction. We developed a novel and simple procedure for constructing multiple alignments based on the profile-profile alignment method and applied it for regular TS targets of CASP10 and CASP ROLL.

Methods

To improve alignment accuracy, we used multiple alignments of a target sequence, homologous sequences of the target, and sequences of template(s). A multiple alignment is obtained by considering minimum inconsistency among all pairwise alignments produced by our profile-profile alignment method, FORTE, with sequences mentioned above.

In addition to the original global-local algorithm of FORTE, we also developed a local profile-profile alignment method and employed it in some cases, especially for 'hard' targets.

Results

As of September 28, 2012, the average TM-score[3]= 0.528 of our method for available 26 domain structures of 20 valid targets in CASP 10 is significantly higher than the one, 0.464, of all methods including modern and sophisticated meta-servers and assembling methods, when we focus on TS1 models of server predictions (=Tarballs) to briefly assess our prediction results.

In the CASP ROLL targets, we could correctly recognize the middle beta-sheet of bacteriophage tail fibers (R0001).

Of course, multiple template-based modeling is effective for building accurate 3D-models. Therefore, we believe that our new method is useful for protein structure prediction.

Availability

The FORTE server is available at http://www.cbrc.jp/forte/[2]. Users can retrieve prediction results as both text and HTML.

1. Tomii,K., Hirokawa,T. & Motono,C. (2005). Protein structure prediction using a variety of profile libraries and 3D verification. *Proteins.* **61**(S7), 114-121.
2. Tomii,K. & Akiyama,Y. (2004) FORTE: a profile-profile comparison tool for protein fold recognition. *Bioinformatics* **20**, 594-595.
3. Xu,J. & Zhang,Y. (2010) How significant is a protein structure similarity with TM-score = 0.5?. *Bioinformatics* **26**, 889-895.

Combining MQAP approaches for improved accuracy in model quality assessment

M. J. Skwark1,2 and A. Elofsson[1,2]

1 – Department of Biochemistry and Biophysics, Stockholm University,
2 – Science for Life Laboratory, Stockholm
arne@bioinfo.se

Pcons and **PconsQ** are model quality assessment methods combining structural consensus (Pcons[1]), and a single model machine learning-based MQAP (ProQ2[2]), as well as – in case of **PconsQ –** distance-based consensus (**PconsD**).

Methods

Structural consensus performs well in terms of ranking protein models, especially given a set of predicted models, which is abundant in • gcorrect • h predictions. **Pcons** (Pcomb in CASP8/9) combines structural consensus with an empirical function scoring the objective quality of single models (ProQ2). Doing so improves the selectivity properties by allowing to discriminate accurately on both ends of quality spectrum, where consensus is less effective: very easy targets (most models are very close to each other) and difficult targets (no meaningful consensus in the model ensemble).

On top of the scoring methods used by **Pcons**, the novel method **PconsQ** employs an additional distance-based metric (**PconsD). Use of inter-residue distance matrices allows a greater emphasis on correct scoring of well-defined structural elements (e.g. structural domains, rigid supersecondary structures etc.) and less on the flexible loop regions.** Models submitted on behalf of **Pcons** and **PconsQ** groups were selected from the set of all predictions submitted in server category in CASP10. The model ensembles obtained from **PconsQ** has been augmented by addition of 10 highest ranked **PconsM** and **PconsD** models each.

The highest ranked models were submitted as predictions for respective methods, without any further refinement.

1. Larsson P., Skwark MJ, Wallner B and Elofssson A (2009) Assessment of global and local model quality in CASP8 using Pcons and ProQ2 *Proteins* **77** (S9): 167-172
2. Ray A., Lindahl E. and Wallner B. (2012) Improved model quality assessment using ProQ2 *BMC Bioinformatics* **13**, 224-
3. McGuffin LJ, Roche D. (2010) Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments *Bioinformatics* **(26)** 182-188

Distance-based modeling by PconsD

M. J. Skwark1,2 and A. Elofsson[1,2]

[1] – *Department of Biochemistry and Biophysics, Stockholm University,* [2] – *Science for Life Laboratory, Stockholm*

arne@bioinfo.se

PconsD is a meta-prediction method, aiming at utilizing consensus inter-residue distances for identification and construction of accurate protein models.

Methods

The method is a based on the set of models produced by **Pcons-net**[1], albeit without rebuilding of unaligned regions. Models are scored by a distance-based model quality assessment approach (see below), which attempts to identify the most correct models in the ensemble as well as the regions of those models in need of rebuilding. Ten best ranked models are subject to rebuilding of identified regions and the resulting ensemble is further rescored by a linear combination of structural consensus (Pcons[2]), distance-based consensus (PconsD) and a single model machine learning-based MQAP (ProQ2[3]) – see abstract for **PconsQ** as MQAP method. **PconsD** as a quality assessment method is based on comparison of inter-residue distance matrices, which allows to overcome many limitations of superposition-based consensus methods (e.g. penalizing of multi-domain proteins and overemphasis of loop regions). Additionally, a streaming algorithm used in implementing the method makes it significantly faster than other MQAP methods available.

Results

Even though **PconsD** MQAP scores do not correlate very well with superposition based metrics such as GDT-TS, one can observe the increased ability of selecting better models from the ensemble, in comparison to superposition-based MQAPs, such as **Pcons.**

Performance-wise, **PconsD** can achieve a three orders of magnitude speed up in consensus-based model quality assessment in comparison to naïve superposition and approx. 8-fold speed-up in comparison to optimized CPU-based structural superposition (**Pcons**). Additionally, proposed approach is approx. 60 times faster than analogous CPU-based methods (ModFOLDclustQ[4]).

1. Wallner B., Larsson P. and Elofssson A (2007) Pcons.net: protein structure prediction meta server. *Nucleic Acids Res.* **35**, suppl 2, W369-W374.
2. Larsson P., Skwark MJ, Wallner B and Elofssson A (2009) Assessment of global and local model quality in CASP8 using Pcons and ProQ2 *Proteins* **77** (S9): 167-172
3. Ray A., Lindahl E. and Wallner B. (2012) Improved model quality assessment using ProQ2 *BMC Bioinformatics* **13**, 224-
4. McGuffin LJ, Roche D. (2010) Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments *Bioinformatics* **(26)** 182-188

Multiple template modeling by PconsM

M. J. Skwark1,2 and A. Elofsson[1,2]

[1] – *Department of Biochemistry and Biophysics, Stockholm University,*
2 – Science for Life Laboratory, Stockholm
arne@bioinfo.se

PconsM is a redesigned approach to consensus based protein structure prediction, aiming at expansion and diversification of model ensembles by utilizing alternative alignments and multiple templates for model building.

Methods

The method is a based on the same set of input alignments and models as **Pcons-net.** Structural templates of 20 best scoring **Pcons-net** models are then subject to realignment using hhalign from HHsuite package, producing alternative alignments.Additionally, alignments producing best scoring models are combinatorially merged to form prospective alignments for multi-template modeling.

Finally, during model building by MODELLER, we employ comprehensive MD-like optimization to ensure good stereochemical features of the model. Models are scored by a linear combination of structural consensus (Pcons[2]), distance-based consensus (PconsD) and a single model machine learning-based MQAP (ProQ2[3]) – see abstract for **PconsQ** as MQAP method.

Results

According to the internal assessment (http://dany.scilifelab.se/CASP10), as of September 26, **PconsM** group is ranked #5 among all server groups in terms of *f*° GDT-TS for all non-canceled CASP10 targets.

1. Soding J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**, 951-960
2. Wallner B., Larsson P. and Elofssson A (2007) Pcons.net: protein structure prediction meta server. *Nucleic Acids Res.* **35**, suppl 2, W369-W374.
3. Larsson P., Skwark MJ, Wallner B and Elofssson A (2009) Assessment of global and local model quality in CASP8 using Pcons and ProQ2 *Proteins* **77** (S9): 167-172
4. Ray A., Lindahl E. and Wallner B. (2012) Improved model quality assessment using ProQ2 *BMC Bioinformatics* **13**, 224-

pcons.net: Improved pipeline for consensus-based protein structure prediction

M. J. Skwark[1,2] and A. Elofsson[1,2]

1 – Department of Biochemistry and Biophysics, Stockholm University,
2 – Science for Life Laboratory, Stockholm
arne@bioinfo.se

The new iteration of pcons.net consensus-based structure prediction server includes multitude of improvements in relation to the CASP9 version, significantly improving resulting model accuracy.

Methods

The method is a development of pcons.net[1] consensus-based protein structure prediction server and as such relies on input from diverse threading methods, such as FFAS, FUGUE, FORTE, HHpred, nFOLD4, rpsblast, SAM-T02 and SAM-T08. The CASP10 version of pcons.net also use HHsuite 2.0[4,5], a suite of threading methods contained in LOMETS[6] package and relies on Rosetta[7] for ab-initio prediction. Regions of the homology models missing in the alignment are rebuilt by an energy-based method in order to ensure compactness of the model and facilitate first-principles based scoring functions. Finally, models are scored by a linear combination of structural consensus (Pcons[2]) and a single model machine learning-based MQAP (ProQ2[3]) – see abstract for **Pcons** as MQAP method. In MQAP category, quality estimates for **Pcons-net** method are obtained by pure structural consensus **(Pcons[2]).**

Results

According to the internal assessment (http://dany.scilifelab.se/CASP10), as of September 26, **Pcons-net** group is ranked #4 among all server groups in terms of *f*°GDT-TS and related scores for all non-canceled CASP10 targets. Additionally, it is ranked #3 by these criteria for hard targets (i.e. those with median GDT-TS of all server predictions below 0.5).

Availability

The improved pipeline will be incorporated into the default pcons.net pipeline in the near future.

1. Wallner B., Larsson P. and Elofssson A (2007) Pcons.net: protein structure prediction meta server. *Nucleic Acids Res.* **35**, suppl 2, W369-W374.
2. Larsson P., Skwark MJ, Wallner B and Elofssson A (2009) Assessment of global and local model quality in CASP8 using Pcons and ProQ2 *Proteins* **77** (S9): 167-172
3. Ray A., Lindahl E. and Wallner B. (2012) Improved model quality assessment using ProQ2 *BMC Bioinformatics* **13**, 224-
4. Soding J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**, 951-960.
5. Remmert M., Biegert A., Hauser A., and S□oding J. (2012) HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173-175

6. Wu S. and Zhang Y. (2007) LOMETS: A local meta-threading-server for protein structure prediction *Nucleic Acids Res.* **35,** 3375-3382

Simulated protein synthesis and folding with residue-residue distance constraints from templates and sequence in Phyre2

L.A. Kelley[1], I. Filippis[1] and M.J.E. Sternberg[1]

[1] – *Imperial College London*
l.a.kelley@imperial.ac.uk

Phyre2[1] (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 is performed using HMM-HMM alignment. Models based on these alignments are used to derive distance constraints for use in a modified version of our *de novo* folding technique, Poing[2]. Additional distance constraints are included from the program PSICOV[3] when sufficient sequence homologues are available.

Methods

A protein sequence is initially scanned against a 50% non-redundant sequence database (Uniref50) using PSI-Blast[4] followed by secondary structure prediction using PSI-pred[5]. A hidden Markov model of the sequence is generated and scanned against a library of HMMs using the HHsearch 2.0.11 package[6]. 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 single template models with no further refinement or loop modelling. The target sequence is also scanned against a 100% non-redundant sequence database using the jackhmmer module of the HMMER3 package[7]. If a sufficiently large number of homologous sequences (>500) are detected the resulting alignment is processed by PSICOV[3] to predict residue-residue contacts.

Each of these simple template 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. Additional weaker springs are included between residues predicted to be in contact by PSICOV. 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. Finally, the resulting models are clustered and the model with the greatest similarity to all other models in the pool is chosen. This Calpha only Poing model has its backbone reconstructed using Pulchra[8]. This full length model is then combined with the original template-based input models using Modeller v9.10[9]. Finally sidechains are placed using our in-house version of the R3 sidechain placement algorithm[10].

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.
4. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
5. Jones,D.T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* ***292***, 195-202.
6. Söding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**, 951-960.
7. Eddy SR (2011) Accelerated profile HMM searches. *PLoS Comput Biol.* **7**(10): e1002195
8. Rotkiewicz, P., Skolnick, J. (2008). Fast method for reconstruction of full-atom protein models from reduced representations *J. Comp. Chem.* **29**, 1460-1465.
9. N. Eswar, M. A. Marti-Renom, B. Webb, M. S. Madhusudhan, D. Eramian, M. Shen, U. Pieper, A. Sali. (2006) Comparative Protein Structure Modeling With MODELLER. *Current Protocols in Bioinformatics*, John Wiley & Sons, Inc., Supplement 15, 5.6.1-5.6.30
10. Xie,W. and Sahinidis,N.V. (2006) Residue-rotamer-reduction algorithm for the protein side-chain conformation problem. *Bioinformatics* **22**, 188-194.

POODLE-I: Disordered regions predictor combining POODLE series with structural information

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

[1] – *Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, Japan*

[2] – *Meiji Pharmaceutical University, Japan*

poodle@cbrc.jp

POODLE server has participated CASP experiment since CASP7. One of the problems that we address is to estimate the regions which prevent protein from crystallizing. These regions are often corresponding to intrinsically disordered regions. To date, we have updated POODLE programs several times. The successful upgrade version, which is called POODLE-I (where "I" stands for integration), employs the workflow approach that was allowed to combine POODLE series with structural information obtained by several other prediction tools[1]. In this round of CASP experiment, we try to add some information into the previous method.

We briefly introduced the original prediction method based on workflow approach. Based on the hypothesis that the factor causing short disordered regions and long ones might be different, the workflow is divided into two parts. One part predicts long disordered regions including unfolded proteins by using POODLE-L, POODLE-W, and COILS (which is coiled coil region predictor). The other part detects short disordered regions by using POODLE-S and several structural information predictors. In this experiment, two steps were introduced to improve performance. In the former part, the domain defined by CATH[2] or predicted by domain linker prediction[3] was assessed whether it was fully disordered or not. In the latter part, we considered that the signal peptide predicted by SingalP[4] or PrediSi[5] are disordered region, because it is usually removed in solving protein structure.

Availability

All POODLE services are available at http://mbs.cbrc.jp/poodle.

1. Hirose,S., Shimizu,K. & Noguchi,T. (2010). POODLE-I: Disordered region prediction by intergrating POODLE series and structural information predictors based on a workflow approach. *In Silico Biol.* **10**, 185-91.
2. Pearl,F.M., Bennett,C.F., Bray,J.E., Harrison,A.P., Martin,N., Shepherd,A., Sillitoe,I., Thornton,J. & Ogengo,C.A. (2003). The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic Acids Res.* **31**, 452-455.
3. Ebina,T., Toh,H. & Kuroda,Y. (2009). Loop-length-dependent SVM prediction of domain linkers for high-throughtput structural proteomics. *Biopolymers.* **92**, 1-8.
4. Petersen,T.N., Brunak,S., von Heijine,G.. & Nielsen,H. (2011) .SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods.* **8**, 785-786.
5. Hiller,K., Grote,A., Scheer,M., Munch,R. & Jahn,D. (2004). PrediSi: prediction of signal peptides and their cleavage positions. *Nucleic Acids Res.* **32**, 375-379.

Prediction of protein disordered region based on conditional neural fields

Masanori. Kakuta[1], Takashi. Ishida[1] and Yutaka. Akiyama[1]
[1] – *Department of Computer Science, Tokyo Institute of Technology*
kakuta@bi.cs.titech.ac.jp

"prdos-CNF" server is an automated protein disordered region prediction server. The predictor was trained by using similar datasets and input vector construction method used in PrDOS method[1] but employs conditional neural fields (CNFs) [2] instead of support vector machines.

Methods

We used first-order (linear-chain) conditional neural fields to predict intrinsically disordered regions. CNF is a supervised machine learning algorithm and an undirected graphical model similar to conditional random field (CRF). CNF can deal with not only dependencies among labels like CRF, but non-linear relationship between input and output. The marginal posterior probability for each position was used as confidence values.
For constructing training sets, we used a non-redundant protein chain set from the PDB using the PISCES server and defined the disordered residues based on the REMARK 465 lines. The input vector for a residue in target sequence is composed of the amino acid types and PSSMs of the sequence in a 27-residue window centered at the residue and the length of the sequence.

1. Ishida. T and Kinoshita K. (2007). PrDOS: prediction of disordered protein regions from amino acid sequence., *Nucleic Acids Res*, **35**, W460-464, 2007.
2. Peng, J. *et al.*, (2009) Conditional neural fields. *Advances in Neural Information Processing Systems (NIPS)*, 1419-1427

Improved model quality assessment using ProQ2

A. Ray[1], E. Lindahl[1] and B. Wallner[2]

[1] *Department of Theoretical Physics & Swedish eScience Research Center, Royal Institute of Technology, Stockholm, Sweden.*

[2]*Department of Physics, Chemistry and Biology & Swedish eScience Research Center, Linköping University, SE-581 83 Linköping, Sweden*
bjornw@ifm.liu.se

Employing methods to assess the quality of modeled protein structures is now standard practice in bioinformatics. In a broad sense, the techniques can be divided into methods relying on consensus prediction on the one hand, and *single-model* methods on the other. Consensus methods frequently perform very well when there is a clear consensus, but this is not always the case. In particular, they frequently fail in selecting the best possible model in the hard cases (lacking consensus) or in the easy cases where models are very similar. In contrast, single-model methods do not suffer from these drawbacks and could potentially be applied on any protein of interest to assess quality or as a direct scoring function for sampling-based refinement.

Methods

ProQ2[1] is an improved *single-model* quality assessment program, based on ideas from its predecessor ProQ[2]. It uses support vector machines to predict local as well as global quality of protein models based on structural and predicted features. All features are calculated over sequence window to achieve a localized prediction. As target function it uses S-score[3]: $S_i=1/(1+(d_i/3)^2)$, where d_i is the distance for residue *i* between the native structure and model, based on a superposition and maximize the sum of S_i. In short, the structural features are similar to the ones used in ProQ: atom-atom contacts, residue-residue contacts, and solvent accessible surfaces. The contacts are encoded as fraction of contacts between 13 different atom types (<4Å for atom contacts), between 6 residue types (<6Å for residue contacts) and as exposure distributions for the same 6 residue types in four exposures bins (<25%, 25%-50%, 50%-75%, and >75%).

In addition to the structural features, predicted secondary structure by PSIPRED[4] and surfaces area by ACCpro[5] was also included. For the secondary structure, three sets of features were calculated: (i) the predicted probability from PSIPRED for the secondary structure of the central residue in the sequence window. (ii) correspondence between predicted and actual secondary structure over a 21-residue window, and (iii) secondary structure assigned by STRIDE[6], binary encoded into three classes over a 5-residue window. For the surface area the correspondence between predicted and actual burial/exposure class over a 21-residue window was used.

Evolutionary information was also included, both directly using sequence conservation calculated from a PSSM and as weighting of the residue based structural features according to the PSSM. For instance, if a position in the sequence profile contains 40% alanine and 60% serine, contacts to this position are weighted by 40% as contacts alanine and by 60% as contacts to serine. This effectively increases the amount of training examples and should also make the final predictor less sensitive to small sequence changes, since data is extracted from multiple

sequence alignments among homologous sequences.

All features described above are localized to short window in sequence to enable a localized prediction. However, it turned out that including the overall correspondence between predicted and actual secondary structure and residue exposure calculated over the whole model instead of a window, improved the performance even for local quality prediction.

In CASP10 ProQ2 participated in the MQAP category and the manual TS category, submitting the highest-ranking server models. It was also linearly combined with Pcons[7] in the Pcomb method (0.8Pcons+0.2ProQ2) and used in a weighted clustering in the ProQ2clust method.

Results

ProQ2 is significantly better than its predecessors at detecting high quality models, improving the sum of Z-scores for the selected first-ranked models by 20% and 32% compared to the second-best single-model method in CASP8 and CASP9, respectively. The absolute quality assessment of the models at both local and global level is also improved. The Pearson's correlation between the correct and local predicted score is improved from 0.59 to 0.70 on CASP8 and from 0.62 to 0.68 on CASP9; for global score to the correct GDT_TS from 0.75 to 0.80 and from 0.77 to 0.80 again compared to the second-best single methods in CASP8 and CASP9, respectively.

Availability

The method is available as a server and standalone download from http://proq2.wallnerlab.org.

1. Ray, A., Lindahl, E. and Wallner, B. (2012). Improved model quality assessment using ProQ2. *BMC Bioinformatics.* **13**(1):224.
2. Wallner, B. and Elofsson, A. (2003) Can correct protein models be identified? *Protein Sci* **12** (5) : 1073-1086.
3. Levitt M, Gerstein M. (1998). A unified statistical framework for sequence comparison and structure comparison. PNAS. 26;95(11):5913-20.
4. Jones, D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* ***292***, 195-202
5. Cheng J, Randall AZ, Sweredoski MJ, Baldi P. (2005). SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Res*, **33**(Web Server issue):W72–W76.
6. Frishman D, Argos P: (1995). Knowledge-based protein secondary structure assignment. *Proteins*, **23**(4):566–579.
7. Wallner, B. and Elofsson, A. (2005) Pcons5: combining consensus, structural evaluation and fold recognition scores. *Bioinformatics* **21** (23) : 4248-4254.

Protein threading by maximizing a new alignment potential

Jinbo Xu[1], Jianzhu Ma[1] and Sheng Wang[1]

[1] –*Toyota Technological Institute at Chicago*
jinboxu@gmail.com

RaptorX consists of the following major components: single-template protein threading, alignment quality assessment (and template selection) and multiple-template threading. Compared to CASP9, we have a new probabilistic graphical model for single-template threading[1]. Since the alignment quality assessment and multiple-template threading methods are almost same as those used in CASP9[2,3], we only present our new single-template protein threading method. In addition, we have also registered another server RaptorX-ZY to test our new method for building 3D models from alignments.

Methods

Given a protein sequence *S* and a template *T*, let *P(A|S,T)* denote the estimated probability of the alignment *A* for two proteins *S* and *T*. The alignment potential of *A* then can be calculated as $\log \frac{P(A|S,T)}{P(A)}$ where *P(A)* is the background (or reference) probability of *A*. In theory, *P(A)* can be calculated by $\Sigma_{X,Y}P(X,Y)P(A|X,Y)$ where *X* and *Y* represents two proteins (with the same lengths as *S* and *T*, respectively) and the summation is calculated over all the protein pairs. In practice, *P(A)* can be approximated by uniformly sampling a few thousand protein pairs. We assume that an alignment is the optimal if it maximizes the alignment potential. That is, to find the optimal alignment between *S* and *T*, we want to maximize $\log \frac{P(A|S,T)}{P(A)}$.

To calculate *P(A|S,T)*, we expand *A* as $\{a_1,a_2,\ldots,a_L\}$ where *L* is the alignment length and a_i is the alignment state at position *i*. In total there are three possible alignment states *M*, I_t and I_s. Meanwhile, *M* represents two residues being aligned, I_t denotes an insertion in the template, and I_s denotes an insertion in the sequence. We use a recently-developed probabilistic graphical model Conditional Neural Field to calculate *P(A|S,T)* as follows[1].

$$P(A \mid T, S, \theta) = \exp\left(\sum_{i=1}^{L} E(a_{i-1}, a_i, T, S) / Z(T, S)\right) \qquad (1)$$

where θ is the model parameter vector to be trained and *Z(T,S)* is the normalization factor (i.e., partition function) summing over all possible alignments for a given protein pair. The function *E* in Eq. (1) estimates the log-likelihood of alignment state transition from a_{i-1} to a_i based upon protein features. We use neural networks to construct the function *E*. The model parameter vector θ consists of all the parameters of the 9 neural networks, which are trained by a set of non-redundant sequence-template pairs. The reference alignments used to train the parameter are generated by our in-house protein structure alignment tool DeepAlign.

We construct the function *E* such that RaptorX has the following properties.

(1) RaptorX explicitly accounts for correlations among protein features by using a nonlinear scoring function (i.e., neural network) to combine a variety of sequence and structure information.

(2) When a protein has a sparse sequence profile[4], RaptorX relies more on structural information since the sequence profile does not contain sufficient information; otherwise it relies more on information in a sequence profile. The structure information includes the 3-class and 8-class secondary structure, the 3-class solvent accessibility and also the structure environment.
(3) RaptorX uses neighborhood information to estimate how likely two residues shall be aligned. The neighborhood information includes sequence profile, secondary structure and solvent accessibility in a small window (size 11) centered at the residues to be aligned. Neighborhood information is especially useful to the weakly similar regions and gap opening positions.
(4) For the disordered regions, RaptorX uses only sequence information since structure information is unreliable. For non-disordered regions, RaptorX uses both sequence and structure information.
(5) Unlike many other methods that use an affine gap penalty, RaptorX uses both position-specific and context-specific gap penalty. The position-specific gap penalty is derived from the alignment of the sequence homologs of a given protein while the context-specific penalty is based upon amino acid identity, hydropathy index, secondary structure and solvent accessibility. When a protein has a sparse sequence profile, RaptorX relies more on context-specific gap penalty; otherwise on the position-specific penalty.

3D model building. By default**,** RaptorX uses MODELLER to build a 3D model from an alignment. We have also developed a new method for model building, which is tested in another CASP10 server RaptorX-ZY. RapotrX-ZY uses a machine learning method to predict distance restraints from an alignment and then build the corresponding 3D model based upon the predicted restraints. The machine learning method predicts distance restraints using information in an alignment including profile similarity and structure similarity.

Results

The CASP10 result is unavailable yet. Here we only present our own test results of the new single-template threading method. In terms of ref-dependent alignment accuracy RaptorX is >10% better than the best profile method HHpred regardless of the benchmarks (see Table 1). To evaluate the quality of the resulting 3D models, given a protein pair we build a 3D model using MODELLER for the target protein based upon its alignment to the template. As shown in Table 2, RaptorX obtains much better 3D models than HHpred, MUSTER and BThreader (i.e., old RaptorX) regardless of the benchmarks, outperforming HHpred by 7-20%.

Table 1. Reference-dependent alignment accuracy on the MUSTER benchmark. Columns 2-5 indicate four different tools generating the reference alignments. Column "BR" indicates the reference alignments provided in the benchmark. Bold indicates the best performance.

Methods	TMalign	Dali	Matt	DeepAlign	BR
HHpred(Local)	42.96	57.34	46.00	46.50	45.34
HHpred(Global)	48.82	53.13	51.48	52.48	51.48
MUSTER	-	-	-	-	46.70
Old RaptorX	47.35	51.30	50.13	50.53	50.01
RaptorX	**54.17**	**58.46**	**57.26**	**59.14**	**57.06**

Table 2. The accumulative model quality, measured by TMscore, on the four benchmarks: In-House, MUSTER, SALIGN and ProSup. Bold indicates the best performance.

Methods	In-House	MUSTER	SALIGN	ProSup
HHpred	1522.77	142.00	121.83	56.44
MUSTER	-	136.47	-	-
BThreader	1537.89	143.95	132.85	66.77
RaptorX	**1692.17**	**152.14**	**134.50**	**67.34**

To further evaluate the modeling performance, we use RaptorX and HHpred to predict the 3D structure for a set of 1000 target proteins randomly chosen from PDB25. All the ~6000 proteins in PDB25 are used as the templates. As shown in Figure 1, RaptorX outperforms HHpred when the target protein does not have a close template. One point in the figure represents two models of a single target. One is built by HHpred and the other by RaptorX. A point above the diagonal line indicates that RaptorX generates a better 3D model. The targets with HHpred TMscore<0.4 usually have sparse sequnece profiles and thus HHpred does not work well for them. By contrast, RapotrX can generate better 3D models for many of them.

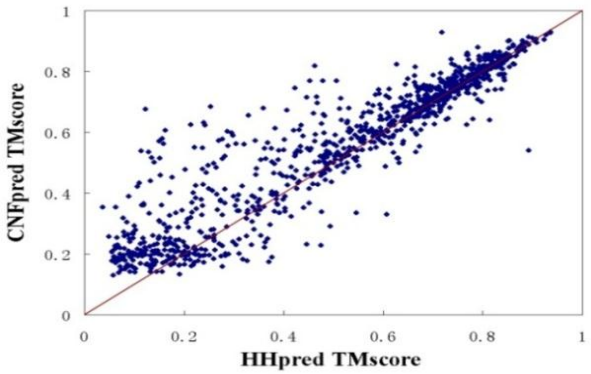


Figure 1. Comparison of RaptorX with HHpred on 1000 proteins.

Availability

The RaptorX server is available at http://raptorx.uchicago.edu. RaptorX currently is running an old protein threading method and we are upgrading it to the new method described in this abstract.

1. Jianzhu Ma, Jian Peng, Sheng Wang and Jinbo Xu. A Conditional Neural Fields model for protein threading. Bioinformatics (Proceedings of ISMB), 2012.
2. Jian Peng and Jinbo Xu. RaptorX: exploiting structure information for protein alignment by statistical inference. PROTEINS, 2011.
3. Jian Peng and Jinbo Xu. A multiple-template approach to protein threading. PROTEINS, 2011.
4. Jian Peng and Jinbo Xu. Low-homology protein threading. Bioinformatics (Proceedings of ISMB 2010), 2010.

Identifying Native-Like Substructure in Protein Decoys – Contact Prediction and Tertiary Structure Prediction

Michael Schneider and Oliver Brock

Robotics and Biology Laboratory, School of Electrical Engineering and Computer Science, Technische Universität Berlin, Einsteinufer 17-EN 10, 10587 Berlin, Germany
oliver.brock@tu-berlin.de

Decoys generated during *ab initio* structure prediction often contain native-like substructures, even in early stages of search. This is to be expected, as native-like substructures represent local, energetically favourable spatial arrangements of amino acids. The ability to differentiate between native-like and "wrong" substructures in decoys is highly relevant to contact prediction and tertiary structure prediction.

We present a machine learning method to identify native-like substructures from an ensemble of decoys. Our method contributes to three of our servers, all described in this abstract. The contact prediction server RBO-CON uses the method to predict contacts from an ensemble of decoys. To do so, it leverages knowledge of favourable physicochemical interactions and occurrence statistics of decoy contacts. In contrast to most existing contact prediction methods, our method does not require multiple sequence-alignments. Instead, it predicts contacts solely based on the target sequence and an energy function.

The tertiary structure prediction servers RBO-i-MBS and RBO-i-MBS-BB leverage native-like substructures of decoys to guide search. In an iterative process (hence the "i" in the server names), the prediction methods generate decoys, identify native-like substructures, extract contact restraints from those substructures, and finally use those restraints in the next iteration of search.

RBO-CON

For contact prediction, we seek to learn the relevant properties that identify native-like contacts from an ensemble of protein decoys. To learn these properties, we specify features that capture physicochemical properties and occurrence statistics of the amino acids in contact. Those features include secondary structure, solvent accessibility, chemical properties of the contacting amino acids, distance distribution and occurrence frequency. We then use a support vector machine (SVM) to learn and predict contacts based on these features.

To devise the training set, we randomly select 400 small proteins (shorter than 100 amino acids) from PDBSelect[1]. We generate decoys for those proteins with our previously developed structure prediction algorithm model-based search (MBS)[2]. Features are then built for the ensembles of the resulting decoys, using only the best 1% decoys ranked by Rosetta's all-atom energy function[3]. The final training set contains approximately 300,000 contacts.

In the learning step, we use an ensemble of SVM classifiers with a bootstrap aggregation[4] (bagging) scheme. Each SVM is trained with a balanced set of 8000 contacts. Parameters for each SVM are tuned to obtain an accurate, but high-variance classifier. The final prediction output is generated by a simple voting scheme of the SVM ensemble.

For each contact prediction, we first perform a structure prediction run with MBS to generate decoys for the target protein. Then, features are computed from the 1% lowest-energy

decoys and predictions are made using the ensemble of SVM classifiers. Predicted contacts are ranked by their number of votes from the individual classifiers.

RBO-i-MBS

The ability to identify native-like substructures in decoys also has benefits in tertiary structure prediction. The tertiary structure prediction server RBO-i-MBS uses the contact predictions generated with RBO-CON as constraints for conformational space search. Each round of model-based search (MBS) generates a set of decoys. Within those decoys, RBO-CON identifies native-like substructures. The contacts contained in these substructures represent constraints for the next round of MBS. Thus, the algorithm iterates between structure and contact prediction to guide search towards regions of the conformational space likely to contain the native state.

RBO-i-MBS-BB

The same iterative scheme of alternating contract prediction and structure prediction is employed in our server RBO-i-MBS-BB. In contrast to RBO-i-MBS, however, it uses a structure prediction method based on structural building blocks. These building blocks are reoccurring, spatially contiguous, sequence non-contiguous structural elements extracted from the PDB, forming a "vocabulary" of protein structure. (For a more detailed description of RBO-MBS-BB, please see the corresponding abstract in this book.)

Availability

A webserver for contact prediction is under development.

1. Griep, S. & Hobohm, U. (2010). PDBselect 1992-2009 and PDBfilter-select. *Nucleic Acids Res* **38**, D318-319.
2. Brunette, T.J. & Brock, O. (2008). Guiding conformation space search with an all-atom energy potential. *Proteins* **73**, 958-972.
3. Rohl, C.A., Strauss, C.E.M., Misura, K.M.S. & Baker, D. (2004). Protein structure prediction using Rosetta. *Meth. Enzymol* **383**, 66-93.
4. Breiman, L. (1996). Bagging predictors. *Machine Learning* **24**, 123-140.

De Novo Structure Prediction Using Model-Based Search

Michael Schneider and Oliver Brock

Robotics and Biology Laboratory, School of Electrical Engineering and Computer Science, Technische Universität Berlin, Einsteinufer 17-EN 10, 10587 Berlin, Germany
oliver.brock@tu-berlin.de

Model-based search (MBS) guides the exploration of conformational space based on information contained in intermediate decoys generated during search. The main algorithmic features of MBS are 1) characterisation of regions as funnels in the energy landscape, 2) accessing the quality of these regions by an all-atom energy function and 3) coordination of computational resources based on this assessment.

The structure prediction method running on our server RBO-MBS has been used in previous CASP experiments (CASP8 and slightly modified in CASP9, formerly known as RBO-Proteus). In CASP10, we used a new implementation of the server used in CASP8. MBS is the search protocol also underlying our other tertiary structure prediction servers in CASP10 (RBO-MBS-BB, RBO-CON, RBO-i-MBS, RBO-i-MBS-BB).

Methods

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

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

Due to our integration with Rosetta[2], model-based search inherits the following algorithmic features. Local search for low-energy conformations starts from an extended backbone conformation. The local, Metropolis Monte Carlo-based search progresses in a number of stages. As the search progresses through the different stages, the move set changes, the number of local search steps are varied, and the accuracy of the energy function is increased. The

energy function progresses gradually from a coarse-grained low-resolution energy function that considers secondary structure, residue environment, and inter-residue pairing to a full-atom energy function that includes side chains and solvation effects.

Each iteration of model-based search uses the same move set and energy function as the corresponding stage in Rosetta. The first stage of model-based search starts after initial 4,000 Monte Carlo fragment insertions have been attempted for each sample. The remaining 32,000 Monte Carlo steps inside Rosetta are divided into the 13 stages of Rosetta's Monte Carlo-based search. For these stages, the parameters of model-based search are adjusted so that each run finishes in approximately 12 hours on 200 processors. For example, proteins with less than 200 residues use 2,000 extended proteins and five all-atom evaluations to evaluate a region. Proteins larger than 200 residues use 1600 extended structures; proteins longer than 300 residues use 100 extended structures. Finally, proteins longer than 500 residues use 600 extended structures. The five lowest scoring models are submitted.

Availability

The code to integrate MBS into Rosetta v2.3 or Rosetta v3.4 is available from the authors on request. A webserver is under development.

1. Brunette, T.J. & Brock, O. (2008). Guiding conformation space search with an all-atom energy potential. *Proteins* 73, 958-972.
2. Rohl, C.A., Strauss, C.E.M., Misura, K.M.S. & Baker, D. (2004). Protein structure prediction using Rosetta. *Meth. Enzymol* 383, 66-93.

Going Beyond Fragments – Using Building Blocks to Guide Protein Structure Prediction

M. Schneider, I. Putz, S. Doerr, F. Salomon, M.Mabrouk, F. Kamm, and O. Brock

Robotics and Biology Laboratory, School of Electrical Engineering and Computer Science, Technische Universität Berlin, Einsteinufer 17-EN 10, 10587 Berlin, Germany
oliver.brock@tu-berlin.de

The PDB is believed to be structurally complete. We therefore mine the PDB for a "vocabulary" of naturally occurring substructures and use this vocabulary for protein structure prediction. Our vocabulary consists of building blocks: spatially contiguous but not necessarily sequence-contiguous structural motives that are repeated in the PDB. We present a method to extract conserved building blocks from the PDB. To account for the fact that structure is more preserved than sequence, this method initially ignores sequence and exclusively operates in the structural domain. Once we have identified building blocks and all their occurrences in the PDB, we can build profiles of the corresponding sequences. These profiles capture part of the sequence space that folds into this unique structure and can later be used for retrieving candidate building blocks for a target sequence. Distance constraints derived from these conserved structural blueprints are used to guide search towards decoys that fulfill the given restraints. Structure predictions are made using our previously introduced search technique model-based search (MBS)[1].

Methods

Our method consists of two stages – one preprocessing stage and one prediction stage. In the offline preprocessing stage, we select a non-redundant set of proteins spanning the fold space of the PDB using the ASTRAL release 1.75[2]. Since we are looking for structurally conserved building blocks, they must occur in more than one protein within this dataset. Thus, we detect partial structural matches between all pairs of proteins to identify recurring structural units using a modified version of Protein3Dfit[3]. These matches consist of several secondary structure fragments, excluding loop regions. Finally, we reduce redundancy through several clustering steps.

Later, in the prediction stage, we must retrieve building blocks based on the target sequence. In order to increase the sensitivity of retrieval, we enhance the sequences associated with each building block with homologous sequences. We generate HMM profiles with HHblits[4] by matching building block fragments against the Uniprot20 sequence database[5]. We now can associate every building block with a set of HMM fragment profiles.

Using the HMM fragment profiles, we can now retrieve relevant building blocks to predict the structure of a target sequence. We first create an HMM profile for the target sequence based on a scan with HHblits against the Uniprot20 database. Then, a matching method based on HHsearch[6] is used to align building block profiles with the target profile. This method allows for arbitrary orderings of building block fragments on the target sequence. The specificity of our profile/profile matching is increased with a technique based on statistical feature analysis over the whole building block database. The resulting retrieval procedure allows us to reliably identify the most relevant building block candidates for a specific target sequence.

To predict the target structure, we constrain conformational space search using the long-range contacts contained in retrieved building blocks. As a result, RBO-MBS-BB focuses search

on regions of the conformational space that favor the partial topology captured by the building blocks.

Availability

A webserver providing access to spatial information guided MBS structure predictions is under development.

1. Brunette, TJ and Brock, O. (2008). Guiding conformation space search with an all-atom energy potential. Proteins **73**, 958-972.
2. Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE. (2004). The **ASTRAL** compendium in 2004. *Nucleic Acids Research **32**:*D189-D192.
3. Lessel, U. and Schomburg, D. (1994). Similarities between protein 3-D structures. *Protein Engineering*, vol. **7**, no. **10**, pp. 1175 –1187.
4. Remmert, M., Biegert, A., Hauser, A. and Soding, J. (2011). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Meth*, vol. advance online publication.
5. UniProt Consortium. (2010). The universal protein resource (uniprot) in 2010. *Nucleic Acids Res*, **38** (Database issue):D124-D148.
6. Söding,J. (2005). Protein homology detection by HMM-HMM comparison. Bioinformatics. **21**, 951-960.

Residue-Residue Contact Prediction by Using Coevolution Information

C.-S. Jeong[1], and D. Kim[1]

[1] - *Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology (KAIST)*
csjeong@kaist.ac.kr, kds@kaist.ac.kr

Because a protein three-dimensional structure is well described by the intra-molecular contacts, predicting residue contact from an amino acid sequence would be useful for reconstructing and validating the protein structure model. Although recent studies have shown that the predicted contact information can be used for inferring the atomic structure, current predictive methodology still needs some improvements of prediction accuracy. Here, we incorporate an improved correlated mutation analysis method, and develop an automated web-server for predicting residue contact by combining the sequence analysis information and SVM method. We assess the prediction performance for CASP9 targets with conventional methods.

Methods

To build a predictive model, we collect 514 proteins from PDB frozen before CASP9 experiment period. The non-redundant proteins are selected from high-resolution X-ray structure <2.5 Å with no missing residues by limiting the maximum sequence identity 25%. The contacts are defined by C_β-C_β distance less than 8 Å. For GLY, C_α is used instead of C_β. In addition, the contacts are categorized to long-, medium-, and short-range contacts, according to the sequence linear distance between residue pair. Long-, medium-, and short-range contact residues are defined as >24, 12-23, and 6-11 aa apart residues in contact, respectively.

The training procedure consists of the following steps. First, given a query sequence, the multiple sequence alignment is constructed by using HHblits[1] with the option "-e 0.001 -n 2," and the profile HMM and the correlated mutation scores are calculated from the multiple sequence alignment. Then, features are extracted from the calculated sequence and evolutionary information and SVM model is constructed for each contact category by using LIBSVM [2].

We calculate 863-dimensional feature vector which represents positional, coevolution, separation segment, and whole sequence information. The positional features consist of spacer, PSSM value, profile-HMM transition probability, effective number of aligned sequences, and predicted secondary structure information. The coevolution features consist of MIp Z-score, MIc Z-score, effective number of aligned sequences, pointwise mutual information. The coevolution scores such as MIp, MIc, and pointwise mutual information are calculated by using profile-based joint probability estimate [3]. It has been shown that the use of profile-based joint probability estimate significantly improves coevolution measurement in contact prediction. The separation segment features consist of within-segment evolutionary amino acid composition, within-segment predicted secondary structure composition, and segment length. The whole sequence features consist of overall evolutionary amino acid composition, overall predicted secondary structure composition, and sequence length. To make the SVM training feasible in a given time period, we only use randomly selected 100,000 examples, keeping the positive-to-negative example ratio as 0.5. Additionally, randomly selected 40,506 examples are used for SVM parameter optimization.

We assess the performance of contact prediction for CASP9 dataset. As following the procedure of CASP9 assessment [4], 28 difficult target domains from 22 CASP9 targets are used. As the main challenge in contact prediction is to improve the long-range contact prediction without template information, we compare the long-range contact prediction accuracy with the conventional template-free methods.

Results

We evaluated the accuracies of long-contact prediction at different rank cutoffs for CASP9 difficult target domains. Our method improves conventional methods at every rank cutoff. Specifically, our method increases the average accuracy by 14.3%, 5.3%, and 18.2% at top-L/5, L/10, and 5 cutoffs, respectively. Since the compared conventional methods were revealed as to outperform other methods in CASP9 results, our method would perform comparably to them.

Availability

Our contact prediction server is freely available at http://binfolab12.kaist.ac.kr/conti/.

1. Remmert, M., Biegert, A., Hauser, A. & Söding, J. (2011) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* **9**, 173-175.
2. Chang, C.-C. & Lin C.-J. (2011) LIBSVM : a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, **2**, 27.
3. Jeong, C.-S. & Kim, D. (2012) Reliable and robust detection of coevolving protein residues. *Protein Eng*. accepted.
4. Monastyrskyy, B., Fidelis, K., Tramontano, A. & Kryshtafovych, A. (2011) Evaluation of residue-residue contact predictions in CASP9 *Proteins* **79 Suppl 10**, 119–125.

Protein structure prediction using epicycles of Monte Carlo sampling

D. Shortle

The Johns Hopkins University School of Medicine
Baltimore, MD 21205
dshortl1@jhmi.edu

Methods

The principal focus of the group is energy-based refinement using Monte Carlo sampling on an ensemble of all heavy atom models. The Monte Carlo moves attempted and the criteria for acceptance are cycled through a sequence of alternatives that vary the type/magnitude of the move and the selective pressure applied. With this strategy more diversity can be introduced and maintained in the evolving ensemble and arrest of refinement in local minima delayed.

Full length automated server models were modified by rebuilding a significant fraction of turns using low Ramachandran energy fragments from the PDB. For refinement targets, the template model and 5-10 server models with the lowest RMSD to the template were reworked instead. From an initial population of 250-500, random samples of 50 models were refined through 4 generations and the best 25 were saved, growing a stage1 pool of 500 to 1000 models. The selective pressure included reducing the atom-atom overlap, which was scored over a changing fraction of residues, plus lowering of atom-level and side-chain level interaction energies and solvation energies. All energy terms were scored with statistical potentials. In many instances the CA-distance matrix error to the ensemble averaged matrix for the preceding generation was also included in the selection function.

The pools generated in Stage1 were refined in a similar manner, with manual adjustments made to the selective pressure in an effort to maintain roughly uniform rates of improvement in all energy terms. In some instances, a third stage of refinement was carried out on the pool of models generated in Stage2.

Typically, the final ensemble had side-chain interaction and solvation energies in the wild-type range, but atom-atom energies and solvation were significantly higher than the typical wild type values. In no case did the Ramachandran energy scored over the turn/loop segments come close to wild-type values, suggesting major errors persisted in the backbone geometry in turns/loops. The final ensemble was K-means clustered, and the five cluster central models were submitted, in an order based on manual inspection of the structures plus measures of compactness and quality of atom packing

Results

Results are limited, but for the 8 refinement targets for which PDB structures were available in mid-September, our best model is closer to the correct structure (by Cα-RMSD) than the template in 5 cases and equal in two. TR722, a tetramer of monomers comprise of one long and one short helix, was the only mis-refined target.

sDisPred – simple disorder prediction using trivial features

L.P. Kozlowski[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*

Protein disorder prediction is an important step during elucidating protein function and in the last years became a standard procedure before protein structure determination. A plethora of programs for protein disorder prediction have been designed. Current state-of-the-art methods are based on machine learning and meta approach (consensus of several primary disorder methods).

On the other hand, protein disorder can be considered as flexible part of a protein which is missing in X-ray structure or represented as random ensemble in NMR spectra. This means that the gaps in alignments against structural databases (e.g. PDB[1]) can be considered as potential disordered regions. Apart from that, the opposite is also possible, i.e. the direct search for disorder in DISPROT database[2] grouping all known disorder proteins. Making use of homology should be very useful as ~90 percent of CASP targets are from TMB category.

Moreover, our sDisPred method uses two additional trivial features. The first feature is the secondary structure. Disordered regions contains considerably less helix and beta sheets content. The second feature is the statistical correction for protein termini which are usually disordered.

sDisPred is an experimental predictor which aim is to establish how well disorder prediction can be done using relatively easily accessible features without any sophisticated machine learning. It can be used also as a baseline for other disorder predictors to assess if they are able to go beyond "trivial" prediction level.

Methods

First part of sDisPred pipeline is a negative and positive scan through databases, PDB and DISPROT respectively, in order to find obvious disorder regions with high homology to those which are already known. This part is done by hhblits[3]. If the hits to the templates from the databases are good, the annotation about disorder is assigned to the target sequence. The regions in the sequence which cannot be reliably aligned to PDB or DISROT templates are considered "new" and they are validated by the second part of the pipeline which predicts disorder and secondary structure using several known programs which are easy to obtain and install locally. For disorder prediction sDisPred uses DISOPRED[4], DisEMBL[5], GLOBPLOT[6], RONN[7], IUPred[8], DISpro[9], DISPROT (VSL2)[10], Metadisorder (by Rost)[11] and SPINE-D[12]. For building secondary structure consensus PSIPRED[13], Prof[14], PROTEUS[15], SSpro4[16], SOPRANO, PSSpred, SPINE-X[17], RAPTOR-XSS[17], SPINE[18] and Netsurfp[19] are used.

In the next part of the pipeline, simple statistical correction for 15 terminal residues is made as those residues are usually more disordered. The statistics is based on REMARK465 annotation taken from PDB header records.

In the final stage, the information from homology searches, disorder and secondary structure consensuses are combined into the ultimate prediction. This is achieved by simple majority rule. Disordered regions are those which are predicted to be disordered by majority of

disorder predictors and contain small fraction of secondary regions.

Results

As the 90 percent of CASP targets are from TMB category it is expected that most of the targets can be easily predicted based on the information from homologs only. This should be also true for disorder prediction. The benchmark based on CASP8 and CASP9 targets confirmed this statement.

Availability

The method will be publicly available in the form of web service if it proves to be valuable in terms of disorder prediction in current CASP.

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. Vucetic,S., Obradovic,Z., Vacic,V., Radivojac,P., Peng,K., Iakoucheva,L.M., Cortese,M.S., Lawson,J.D., Brown,C.J., Sikes,J.G., Newton,C.D. & Dunker,A.K. (2004). DisProt: a database of protein disorder. *Bioinformatics*.
3. Remmert,M., Biegert,A., Hauser,A. & Soding,J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* **9**, 173-5.
4. 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.
5. 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.
6. 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.
7. 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.
8. 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.
9. Cheng,J., Sweredoski,M. & Baldi,P. (2005). Accurate prediction of protein disordered regions by mining protein structure data. *Data Mining and Knowledge Discovery* **11**, 213-222.
10. Obradovic,Z., Peng,K., Vucetic,S., Radivojac,P. & Dunker,A.K. (2005). Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins* **61 Suppl 7**, 176-82.
11. Schlessinger,A., Punta,M., Yachdav,G., Kajan,L. & Rost,B. (2009). Improved disorder prediction by combination of orthogonal approaches. *PLoS One* **4**, e4433.
12. Zhang,T., Faraggi,E., Xue,B., Dunker,A.K., Uversky,V.N. & Zhou,Y. (2012). SPINE-D: accurate prediction of short and long disordered regions by a single neural-network based method. *J Biomol Struct Dyn* **29**, 799-813.
13. McGuffin,L.J., Bryson,K. & Jones,D.T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics* **16**, 404-5.
14. Ouali,M. & King,R.D. (2000). Cascaded multiple classifiers for secondary structure

prediction. *Protein Sci* **9**, 1162-76.
15. Montgomerie,S., Cruz,J.A., Shrivastava,S., Arndt,D., Berjanskii,M. & Wishart,D.S. (2008). PROTEUS2: a web server for comprehensive protein structure prediction and structure-based annotation. *Nucleic Acids Res* **36**, W202-9.
16. Pollastri,G., Przybylski,D., Rost,B. & Baldi,P. (2002). Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins* **47**, 228-35.
17. Faraggi,E., Zhang,T., Yang,Y., Kurgan,L. & Zhou,Y. (2012). SPINE X: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. *J Comput Chem* **33**, 259-67.
18. Dor,O. & Zhou,Y. (2007). Real-SPINE: an integrated system of neural networks for real-value prediction of protein structural properties. *Proteins* **68**, 76-81.
19. Petersen,B., Petersen,T.N., Andersen,P., Nielsen,M. & Lundegaard,C. (2009). A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct Biol* **9**, 51.

Protein Model Structure Refinement by Physics-based Relaxation and Loop Modeling

Lim Heo, Hahnbeom Park, Gyu Rie Lee and Chaok Seok

Department of Chemistry, Seoul National University, Seoul 151-747, Republic of Korea
chaok@snu.ac.kr

Protein model refinement methods have had more difficulties in improving global structure quality than in improving local structure quality, according to the assessments of previous CASPs. In CASP9, our refinement method could improve global structure quality on average but showed poor performance in local structure improvement. This result was partly due to the fact that we did not refine the overall model structures but re-modeled loop or terminus regions only. In CASP10, we performed both overall structure relaxation and loop modeling, achieving consistent improvements in both global and local structure accuracies for the targets whose experimental structures have been released so far.

Methods

As shown in the flowchart of **Figure 1**, we tried two refinement methods, a mild refinement method and a more aggressive refinement method. The model generated by the mild refinement method was submitted as model 1, and 4 additional models generated by the aggressive method were submitted as models 2~5. Both methods are based on repeated relaxations of model structures by molecular dynamics simulations after structure perturbations. Subsequent loop modeling was performed only after the mild refinement to save computation time.

The energy functions used for the two relaxation methods are linear combinations of a physics-based energy function complemented by database-derived terms and a restraint energy derived from the given initial model structure. The relative weight of the restraint energy to the physics-based energy for the mild relaxation was five times larger than that for the aggressive relaxation. The physics-based energy function contains molecular-mechanics bonded energy terms, Lennard-Jones potential energy, Coulomb energy, hydrogen bond energy, FACTS solvation energy, solvent accessible surface area energy, dDFIRE potential energy, and sidechain and backbone torsion angle energy.

Structure perturbations were applied only to clusters of side-chains in the mild refinement, and more aggressive perturbations to secondary structure elements and loops were applied in the aggressive refinement. The triaxial loop closure method[1,2] was employed to avoid breaks in model structures caused by perturbations to internal torsion angles.

An *ab initio* protein loop modeling was carried out after the mild relaxation for a maximum of five unreliable loop or terminus regions detected by a method based on ProQres. The loop modeling method searches for the global optimum of another physics-based energy enforced by free energy components from database-derived potentials that was designed for modeling unreliable protein loops and termini[3,4].

Results

Refinement results for the 8 out of 27 refinement targets whose experimental structures have been released so far (Sep. 8, 2012) were analyzed. Changes in GDT-HA, GDC-SC, and

MolProbity scores were calculated for the models before and after refinement to measure improvements in global structure accuracy, local structure accuracy, and physical correctness, respectively. For the structures submitted as the first models, average improvements in GDT-HA, GDC-SC, and MolProbity scores are 1.55, 3.17, and 1.23, respectively. When the best models among the submitted models are considered, our refinement method could improve models for all 8 targets in all 3 measures with average improvements of 3.34 (GDT-HA), 4.22 (GDC-SC), and 1.55 (MolProbity).

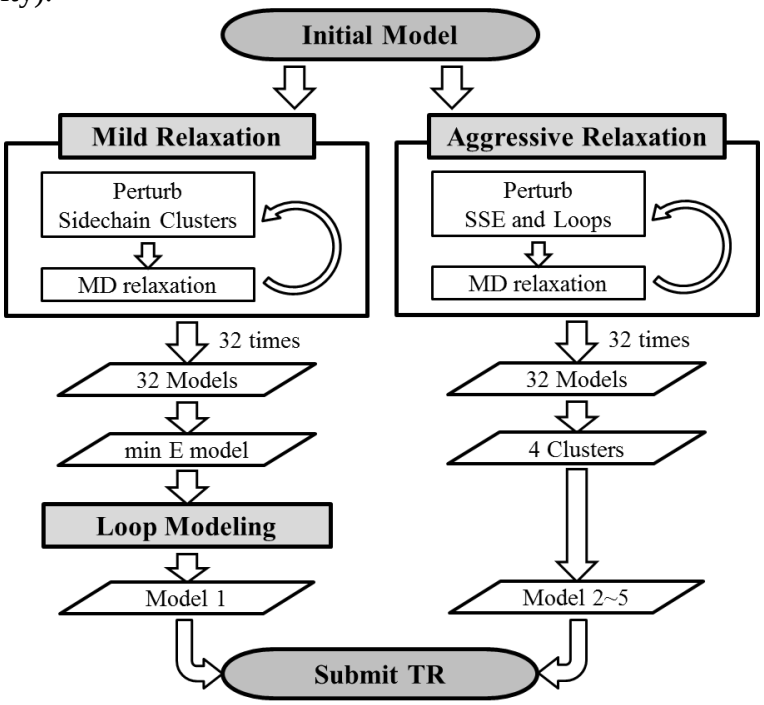


Figure 1. Flowchart of the refinement method tested by the Seok group in CASP10. Two relaxation methods, a mild relaxation and a more aggressive relaxation, and loop modeling were tested.

Availability

A web server for this method will be constructed at http://galaxy.seoklab.org.

1. Coutsias,E.A., Seok,C., Jacobson,M.P. & Dill,K.A. (2004) A kinematic view of loop closure. *J. Comput. Chem.* **25**, 510-528.
2. Lee,J., Lee,D., Park,H., Coutsias,E.A. & Seok,C. (2010) Protein loop modeling by using fragment assembly and analytical loop closure, *Proteins: Structure, Function, and Bioinformatics*, **78**, 3428-3436.
3. Park,H. & Seok,C. (2012) Refinement of unreliable local regions in template-based protein models, *Proteins: Structure, Function, and Bioinformatics*, **80**, 1974-1986.
4. Park,H., Ko,J., Joo,K., Lee,J., Seok,C. & Lee,J. (2011) Refinement of protein termini in template-based modeling using conformational space annealing, *Proteins: Structure, Function, and Bioinformatics*, **79**, 2725-2734.

GALAXY in CASP10: Modeling Reliable Protein Core Regions and Refining Unreliable Regions

Lim Heo, Hahnbeom Park, Junsu Ko, Gyu Rie Lee, Hasup Lee, Woonghee Shin, Minkyung Baek and Chaok Seok

Department of Chemistry, Seoul National University, Seoul 151-747, Republic of Korea
chaok@snu.ac.kr

Seok-server performed fully automated tertiary structure modeling from domain parsing to homo-oligomer prediction in CASP10. It employed several modules of GALAXY biomolecular modeling package such as GalaxyCassiopeia for template-based model building, GalaxyRefine for loop/terminus modeling[1,2], GalaxyPersus for free modeling, GalaxyGemini for homo-oligomer structure prediction (manuscript submitted), and GalaxySite for ligand binding site prediction (manuscript submitted). HHsearch[3] and PROMALS3D[4] were used at the initial stages of multiple template selection and sequence alignment.

Methods

The tertiary structure prediction pipeline for Seok-server is shown in **Figure 1**. For a given query sequence, segments of the sequence for which template-based modeling are possible are first detected by a method that uses HHsearch results. Such segments are called restraint units (RUs) because spatial restraints necessary for template-based model building are derived independently for each RU from the corresponding multiple templates. Confidence for each RU is estimated from the qualities of the selected templates and their alignments. If more than one RU of high confidence is detected, template-based modeling (TBM) is performed. If only RUs of medium confidence exist, both TBM and free modeling (FM) are performed. Otherwise, FM is performed.

In the TBM process, alignment of core sequence with those of templates is obtained by PROMALS3D, and models are built by a new method called GalaxyCassiopeia which replaces MODELLER and MODELLERCSA used by us in the previous CASP. GalaxyCassiopeia first derives spatial restraints for model building from templates, performs quick optimizations to generate multiple models from the restraints to detect unreliable regions and unreliable side-chains. More optimization efforts are put to optimize the unreliable side-chains in the following stage in which the overall structure is relaxed using a physics-based energy combined with the restraints from templates. Long unreliable regions (LURs), which may even correspond to a domain of FM target in some cases, are modeled by a method based on our FM method called GalaxyPerseus. Short unreliable regions (SURs) are modeled by our *ab initio* loop/terminus modeling method. Homo-oligomer structure is also predicted after tertiary structure models are generated by selecting oligomer templates using a new similarity-based method called GalaxyGemini.

GalaxyPerseus is a structure prediction method for those targets lacking reliable template information. In contrast to the other GALAXY methods, GalaxyPerseus uses a coarse-grained molecular representation to reduce the conformational search space. Secondary structure

elements are first sampled by a Monte Carlo search and then assembled by optimizing a physicochemical energy using a genetic algorithm method.

Finally, ligand binding sites are predicted by the GalaxySite method which performs molecular docking simulations for candidate ligands selected by a similarity-based method.

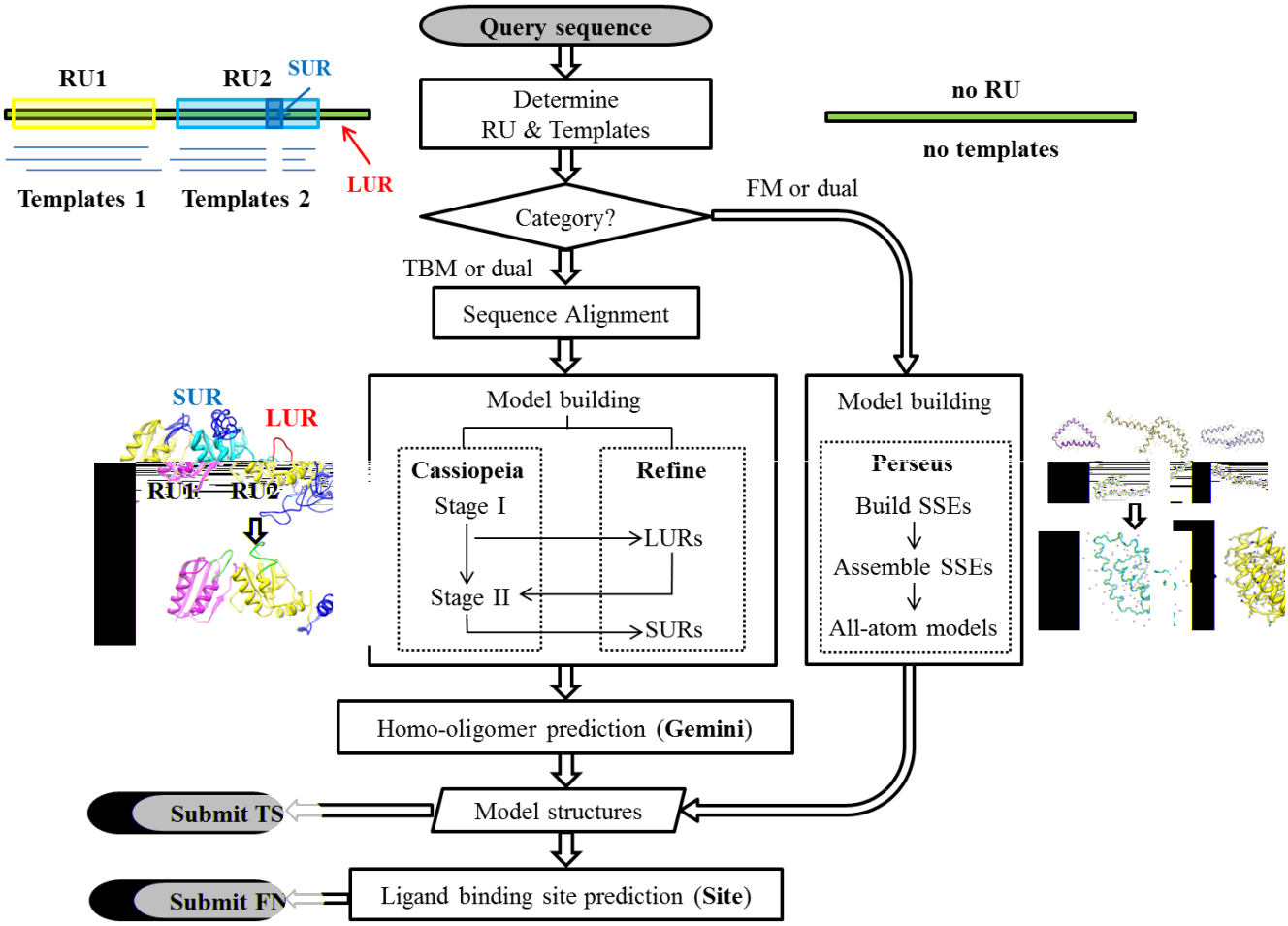


Figure 2. Flowchart of tertiary structure modeling and binding site prediction of Seok-server in CASP10

Availability

Some parts of the above method including an older version of template-based modeling, loop/terminus modeling, homo-oligomer prediction, and binding site prediction are freely available as web servers at http://galaxy.seoklab.org.

1. Park,H. & Seok,C. (2012) Refinement of unreliable local regions in template-based protein models, *Proteins: Structure, Function, and Bioinformatics*, **80**, 1974-1986.
2. Park,H., Ko,J., Joo,K., Lee,J., Seok,C. & Lee,J. (2011) Refinement of protein termini in template-based modeling using conformational space annealing, *Proteins: Structure, Function, and Bioinformatics*, **79**, 2725-2734.

3. Soding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**, 951-960.
4. Pei,J., Kim,B.H. & Grishin,N.V. (2008) PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic acids research*, **36**, 2295-2300.

Slave to the machine

R.B. Sessions

School of Biochemistry, University of Bristol, BS8 1TD, UK
r.sessions@bristol.ac.uk

CASP9 was my first complete "All groups" submission. A ranking around 127 illustrated that traditional homology modelling did not work too well into the grey zone. No surprises there for anyone, but it was a useful personal exercise. In CASP10 I am trying another "dumb" exercise, acting a bit like a minimally intelligent server interface. Hence I have used HMM and automated modelling methods to generate 5 possible models per target, just as I might in a real collaboration with an experimental group requiring a model in the grey zone. The only additional science is that the 5 models were ranked by a combination energy evaluation via our empirical free energy forcefield and visual inspection.

Methods

The HHpred sever was used with default settings to search for appropriate templates. MODELLER was used to generate atomic models from hits. Typically the five models generated comprised: HHpred's best single template; HHpred's best set of multiple templates; three user selected sets of template structures. The models were scored using the BUDE forcefield[1,2]. The final ranking was determined by BUDE score and visual inspection.

Availability

The major tools HHpred and MODELLER are available from the Söding and Sali groups respectively. BUDE is available on request to the author.

1. Gibbs,N., Clarke,A.R., Sessions,R.B. (2001). Ab initio protein structure prediction using physicochemical potentials and a simplified off-lattice model *Proteins*. **43,** 186-202.
2. McIntosh-Smith,S., Wilson,T., Avila-Ibarra,A., Crisp,J., Sessions,R.B. (2011). Benchmarking energy efficiency, power costs and carbon emissions on heterogeneous systems. *The Computer Journal*. **55,** 192-205.

SP-ALIGN server for binding site prediction in CASP10

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 SP-ALIGN server is an update of the FINDSITE[1] approach for binding site prediction.

Method

FINDSITE is a threading-based method that detects binding pockets for small molecules[1]. Protein threading is capable of detecting remote, yet evolutionary related homologues. The conservation of functional sites among homologous proteins allowed us to develop FINDSITE, a highly accurate method for ligand-binding site prediction and functional annotation. FINDSITE employs template identification, structure superimposition and binding site clustering as follows: First, for a given target sequence, structure templates are selected by three threading procedures: PROSPECTOR_3[2], SPARKS[3] and SP[34]. Subsequently, template structures bound to ligands are identified and superimposed onto the target protein structure using the structural alignment algorithm TM-align[5]. Then, the centers of mass of ligands bound to threading templates are clustered according to their spatial proximity, using an 8-Å cutoff distance. This cutoff maximizes the ranking accuracy and accommodates some structural distortions. The geometrical center of each cluster corresponds to the center of a putative binding site. Finally, the predicted binding sites are ranked according to the number of threading templates that share a common binding pocket (cluster multiplicity). For the SP-ALIGN server, we use the models predicted by the latest version of TASSER methodology TASSER^{VMT}-lite[6] as the reference structures and the structural alignment is updated to include a heuristic structure-pocket alignment (SP-ALIGN) to filter the template pockets and to derive the rotational and translational matrix of template ligands. After the superimposition, putative binding sites are inferred through the clustering of the template ligands, and the predicted sites are ranked according to the alignment score of structure-pocket alignment. Benchmarking carried out for the 30 binding site prediction targets of the 9th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP9)(http://predictioncenter.org/casp9/) gives a Matthew's correlation coefficient (MCC) of 0.71 between predicted and observed binding residues. This performance is indistinguishable from the best Human prediction in CASP9.

Availability

The SP-ALIGN web service is available at
http://cssb.biology.gatech.edu/skolnick/webservice/casp/SP-ALIGN/index.html

1. Brylinski M, Skolnick J: FINDSITE: A threading-based method for ligand-binding site prediction and functional annotation. *Proc Natl Acad Science* 2008, **105:**129-134.
2. Skolnick J, Kihara D, Zhang Y: Development and large scale benchmark testing of the PROSPECTOR 3.0 threading algorithm. *Proteins* 2004, **56:**502--518.

3. Zhou H, Zhou Y: Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. *Proteins* 2004, **55** 1005--1013.
4. Zhou H, Zhou Y: Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. *Proteins* 2005, **58** 321--328.
5. Zhang Y, Skolnick J: TM-align: a protein structure alignment algorithm based on the TM-score. *Nucl Acids Res* 2005, **33:**2302--2309.
6. Zhou H, Skolnick J: FINDSITE^{X}: A structure based, small molecule virtual screening approach with application to all identified human GPCRs. *Molecular Pharmaceutics* 2012, **9**(6):1775-1784.

Augmenting Phyre2 with server models and structure searching

L.A. Kelley[1*], I. Filippis[1] and M.J.E. Sternberg[1]

[1] – *Imperial College London*
l.a.kelley@imperial.ac.uk, i.filippis@imperial.ac.uk

Methods

Human 3D structure predictions were made using structural clustering with a modified version of the Poing[1] *de novo* modeling tool as described in the Phyre2[2] CASP10 abstract and structurally searching potential models against the PDB. 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. In targets containing multiple domains, clustering was performed at the domain level, and the highest ranking domains reconnected using Poing.

For very difficult FM targets, between 500 and 10,000 models were produced by Poing and clustered. The centroid model from each cluster was then searched against a representative database of protein structures using MAMMOTH[4]. High scoring structural matches to known structures (MAMMOTH E-value < 10^{-3}) were considered as potentially correct and the structural alignment from MAMMOTH used to adjust the position of backbone atoms in the final model to bring them closer to the matched PDB structure. This process was sometimes iterated 2-3 times depending on the connectivity of the resulting backbone.

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. Ortiz AR, Strauss CE, Olmea O. (2002) MAMMOTH (Matching molecular models obtained from theory): An automated method for model comparison. *Protein Sci.***11**(11):2606-21.

Selection of Templates Recursively by INteGrating exhaustive Strategies (STRINGS) server

G. Chopra[1, 2], H. Tiang[1] and R. Samudrala[1]

1 - Dept. of Microbiology University of Washington USA, 2 - Dept. of Structural Biology Stanford University USA

ram@compbio.washington.edu

STRINGS is an automated protein 3D structure prediction server made by combining several state-of-the-art methods. Specifically, we used a combination of HHSearch[1], I-TASSER v.1.0 standalone package[2; 3] and KoBaMIN standalone package[4; 5] for generating 3D-structures of proteins. Template-based modeling has been the most successful method in recent CASP experiments, however template selection still remains a challenge, as it is not always possible to identify the best template in the PDB library, due to the limitations of the alignment methods.

Methods

We address the problem directly in an exhaustive fashion by recursively splitting the query protein into several segments and then use HHSearch[1] to identify the best template match for each segment. This procedure helps to identify both the domain boundaries, as well as the template with the best alignment for this local region. The final selection of templates and domain boundary is done based on the threading alignment Z-score and alignment coverage. Once the optimal domains are identified, we model them individually using a modified version of I-TASSER v.1.0 standalone package[2; 3], which includes multiple template identification by both LOMETS[6] and HHSearch threading programs. Top five models generated from this approach are refined using the consistent refinement protocol implemented by the KoBaMIN refinement server[4; 5] (http://csb.stanford.edu/kobamin/) and submitted automatically. For cases, where no optimal domain boundaries are identified even after splitting the sequence, the entire sequence is processed as such by the modified version of I-TASSER v.1.0 standalone package and then by KoBaMIN protocol for submission.

Availability

STRINGS is available as a web sever at http://cando.compbio.washington.edu/casp/strings.

1. Soding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics.* **21**, 951-60
2. Roy A., A Kucukural A. & Zhang Y. (2010) I-TASSER: a unified platform for automated protein structure and function prediction. *Nature Protocols*. **5**, 725-738.
3. Zhang Y. (2008) I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*. **9**, 40.
4. Chopra,G., Kalisman,N. & Levitt,M. (2010). Consistent refinement of submitted models at CASP using a knowledge-based potential. *Proteins.* **78**, 2668-2678.
5. Rodrigues, J. P. G. L. M., Levitt, M. & Chopra, G. (2012). KoBaMIN: a knowledge-based minimization web server for protein structure refinement. *Nucleic Acids Res.* **40**, W323-W328.

6. Wu S. & Zhang, Y. (2007) LOMETS: A local meta-threading-server for protein structure prediction. *Nucleic Acids Res* **35**, 3375-3382.

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 $-\mathrm{C_a}-\mathrm{CO}-\mathrm{NH}-$ 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.

Add all-atom side chain to residue

Since crank model can provide atom locations for backbone atoms, the central problem is how to determine the sidechain atom coordinates if the atom coordinates are known for a backbone structure in arbitrary orientation. Thanks for the knowledge of amino acid structure, we have the atom coordinates for sidechain in some special orientation. As a consequence, we can determine the sidechain atom coordinates by matching amino acid to the backbone of a crank.

As the structure of 20 amino acids are well determined by experiment observation, we have the atom coordinates for any type of residues, including backbone $\mathbf{X}_{BB}^{obs}$ and sidechain $\mathbf{X}_{SC}^{obs}$. The only problem is that the observed amino acid structure are usually not in the same orientation as in crank model. If the backbone parts N-Cα-C-O of observed amino acid structure overlap with

crank model (**Error! Reference source not found.**), i.e., $\mathbf{X}_{BB}^{obs} = \mathbf{X}_{BB}^{crank}$, it is obvious that the rank sidechain atom will be determined by $\mathbf{X}_{SC}^{crank} = \mathbf{X}_{SC}^{obs}$. By multiplying a rotation matrix $M^{obs-crank}$, the observed amino acid structure can easily overlap the crank.

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.

(a) 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. A pair of residues are considered in contact if any two non-hydrogen side chain atoms (NHSA) from residues i, j are within a specified cutoff distance. In AA-CSAW, we use the Atom Distance criteria (ADC) model(Sun and He, 2010; Sun and He, 2011) in residue contact determination.

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.

(b) 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.

(c) 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., *et al.* (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*. 2007 Ieee International Conference on Bioinformatics and Biomedicine Workshops, Proceedings.
8. Sun, W. and He, J. (2010) Understanding on the Residue Contact Network Using the Log-Normal Cluster Model and the Multilevel Wheel Diagram, *Biopolymers*, **93**, 904-916.
9. Sun, W.T. and He, J. (2011) From Isotropic to Anisotropic Side Chain Representations: Comparison of Three Models for Residue Contact Estimation, *Plos One*, **6**.

TASSER for protein structure prediction in CASP10

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 Human expert group TASSER has submitted predictions for protein structures and explored a number of refinement protocols for refinement targets.

Method

Our human expert prediction is semi-automated. For a given target, we download all models of all CASP structure prediction servers. Based on the coverage of top models of our threading method SP^3 [1], the target is divided into domains. Then, the downloaded structures are parsed into corresponding domains. Three model quality assessment prediction methods: (a) GOAP[2], (b) FTCOM[3] and (c) TASSER-QA[4] are used to rank the domain structures. Targets are also 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 each ranking method, we select the top 30 ranked structures for further TASSER refinement[5]. Tertiary restraints and contacts are then derived from those selected models. For Medium/Hard targets, additional chunk models from chunk-TASSER[6] are also included to derive restraints and contacts. The distance restraints and contacts are then fed into TASSER to refine the selected models. We performed a single long simulation of TASSER followed by SPICKER[7] clustering for each domain and each ranking method. The top first cluster centroid model is selected from each TASSER simulation. 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 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[8] energy function with a simple sampling procedure by the changing side-chain conformations sequentially along the chain. Final model ranking is based on a benchmarking study of CASP9 targets: For Easy targets, the first model is the refined model from GOAP selection, the second is from FTCOM, the third is from TASSER-QA, the fourth is the top unrefined GOAP selection and the fifth is the top unrefined TASSER-QA selection. For Medium/Hard targets, the first model is the refined model from TASSER-QA selection, the third is from GOAP and the rest are the same as for Easy targets.

We have used two methods for refinement targets. One is loop modeling by generating a large number of alternative loop conformations based on the information provided by CASP organizers if available or secondary structure predictions and using the GOAP energy function to

select top 5 unique models. We then use the 5 models as effective multiple template models in MODELLER[9] to build final model. The other is to select top 30 closest models to the refinement target from all CASP servers and feed them into TASSER for refinement.

Result

The predicted structures by TASSER human expert have better quality than our TASSER-VMT server predictions by about 13% as assessed by their GDT-TS-score to native for the top first models of the released 30 human targets/domains (by Sept. 20, 2012). This is mainly due to the better pool of structures from the CASP servers and reliable GOAP and TASSER-QA model selections. The TASSER prediction is also slightly better than the top server for these 30 targets/domains (total GDT-TS-score 15.19 vs. 14.91).

Availability

TASSER related programs as well as their services are available through our webpage at http://cssb.biology.gatech.edu/

1. 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.
2. Zhou, H. & Skolnick, J. (2011). GOAP: A Generalized Orientation-Dependent, All-Atom Statistical Potential for Protein Structure Prediction. Biophysical Journal **101**, 2043-2052.
3. Zhou,H and Skolnick, J. (2010) Improving threading algorithms for remote homology modeling by combining fragment and template comparisons. Proteins. **78**, 2041-8.
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. Zhang, Y. and J. Skolnick(2004) Automated structure prediction of weakly homologous proteins on genomic scale. Proc. Natl. Acad. Sci. (USA) **101**,7594--7599.
6. Zhou, H and Skolnick, J. (2007) Ab initio protein structure prediction using chunk-TASSER.. Biophysical Journal. **9**3,1510-8.
7. Zhang, Y. and Skolnick,J. (2004) SPICKER: a clustering approach to identify near-native protein fold. J. Comput Chem **25**, 865--871.
8. Zhou, H., and Y. Zhou. 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.
9. Sali,A., et.al. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993;234:779--815.

TASSER-VMT server for protein structure prediction in CASP10

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 TASSER-VMT server implements the latest variants of the TASSER methodology for automated protein structure prediction[1,2].

Method

Based on the observation that multiple template-based methods often perform better than single template-based methods, we explore the use of a **V**ariable number of **M**ultiple **T**emplates (VMT) for a given target in the latest variant of TASSER[2], TASSER[VMT]. We first develop an algorithm that improves the target-template alignment for a given template. The improved alignment, called the SP[3] alternative alignment, is generated by a parametric alignment method coupled with short TASSER refinement on models selected using knowledge-based scores. The refined top model is structurally aligned to the template to produce the SP[3] alternative alignment. Templates identified using SP[3] threading[3] are combined with the SP[3] alternative and HHEARCH alignments to provide target alignments to each template. These template models are then grouped into sets containing a variable number of template/alignment combinations. For each set, we run short TASSER simulations to build full-length models. Then, the models from all sets of templates are pooled, and the top 20-50 models selected using FTCOM[4] ranking method. These models are then subjected to a single longer TASSER refinement run for final prediction. We benchmarked our method by comparison with our previously developed approach, pro-sp3-TASSER[5], on a set with 874 Easy (defined as having SP[3] Z-score >=6) and 318 Hard targets (SP[3] Z-score < 6). The average GDT-TS score improvements for the first model are 3.5% and 4.3% for Easy and Hard targets, respectively. When tested on the 112 CASP9 targets, our method improves the average GDT-TS scores as compared to pro-sp3-TASSER by 8.2% and 9.3% for the 80 Easy and 32 Hard targets, respectively. It also shows slightly better results than the top ranked CASP9 Zhang-Server, QUARK and HHpredA methods.

Result

An in-house assessment of TASSER-VMT's performance for the 56 released targets/domains by Sep. 20, 2012 shows that TASSER-VMT performs among the top servers.

Availability

The TASSER-VMT service is available through our webpage at http://cssb.biology.gatech.edu/

1. Zhou, H, Skolnick, J. (2012) Template-based protein structure modeling using TASSER[VMT]. Proteins: Structure, Function, and Bioinformatics. **80**(2):352-361
2. Zhang, Y. and J. Skolnick(2004) Automated structure prediction of weakly homologous proteins on genomic scale. Proc. Natl. Acad. Sci. (USA) **101**,7594--7599.

3. 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.
4. Zhou,H and Skolnick, J. (2010) Improving threading algorithms for remote homology modeling by combining fragment and template comparisons. Proteins. **78**, 2041-8.
5. Zhou,H and Skolnick, J. (2009) Protein structure prediction by pro-sp3-TASSER. Biophysical Journal. **96**, 2119-27.

Hand Building Predictive Models Using An Amino Acid Structural Code

Archana G. Chavan, Hyun Joo, Jamie Phan, Jerry Tsai

University of the Pacific, Stockton CA 95211

In an exhaustive analysis of packing in the Protein Data Bank, the knob-socket tetrahedral construct was identified as a fundamental principle underlying protein structure (1,2). Application of this knob-socket principle to classification of protein structure reveals distinct amino acid preferences for certain knob-socket arrangements. These preferences define a discrete amino acid code for the relative spatial arrangement of protein residues in secondary and tertiary structure. Amino acid composition of 3-residue sockets specifies secondary structure, while interaction of the 3-residue socket with a fourth residue indicates tertiary packing. Our approach applies this amino acid code to adjust secondary structure predictions and precisely pack these secondary structure elements in an essentially hand building process.

Methods

In this past CASP10 experiment, 54 regular target sequences were modeled: 21 template-based models (TBM) and 33 free models (FM). In our approach, secondary structure is first identified and refined. For TBM targets, secondary structure is obtained from structural template identification using PSI-BLAST (3). For FM targets, an initial secondary structure analysis is performed using PSIPRED(4) and/or FFAS(5). Primary sequence alignment and secondary structure are then modified based on the 3-residue socket propensities. The next step involves the topological assembly of these secondary structure elements. Our knob-socket based amino acid code was used to map out patterns of knob residues packing into particular 3 residue sockets between secondary structure elements. The TBM targets largely involved rearrangements of the template knob-socket patterns, while FM targets required de novo identification of knob-socket patterns. The UCSF Chimera package (6) was used to place secondary structure elements, and the Modeller (7) package was used to build final models. In a similar fashion, predictions were constructed for 5 refinement and 9 assisted modeling targets.

Availability

As this approach based on a knob-socket defined amino acid code is still under development, a defined package of code is yet to be developed. However, the knob-socket propensities are available upon request.

1. Joo,H., Chavan,A.G., Phan,J., Day,R. and Tsai,J. (2012). An amino acid packing code for α-helical structure and protein design. J. Mol. Biol. 419, 234-254.
2. Day,R., Lennox,K.P., Dahl,D.B., Vannucci,M. & Tsai,J.W. (2010). Characterizing the regularity of tetrahedral packing motifs in protein tertiary structure. Bioinformatics 26, 3059–3066.
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. McGuffin,L.J., Bryson,K. and Jones,D.T. (2000). The PSIPRED protein structure prediction

server. Bioinformatics. 16, 404-405.
5. Jaroszewski,L., Rychlewski,L., Li,Z., Li,W. and Godzik,A. (2005). FFAS03: a server for profile-profile sequence alignments. Nucl. Acids Res. 33, W284-W288.
6. Pettersen,E.F., Goddard,T.D., Huang,C.C., Couch,G.S., Greenblatt,D.M., Meng,E.C. and Ferrin,T.E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. J. Comput. Chem. 25, 1605-1612.
7. Sali,A. and Blundell,T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234, 779-815.

Single-model quality assessment based on a distance map prediction

Genki Terashi, Yuuki Nakamura, Hiromitsu Shimoyama and Mayuko Takeda-Shitaka

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

In the previous CASP experiments, the consensus based method has been shown to outperform other method for Quality Assessment (QA) category. The top performed QA groups used very similar strategies (consensus method) and obtained very similar results. On the other hands, according to the correlation-based assessment in CASP9, single model methods were far behind the consensus methods and it indicated that there is room to improve the single model method. Therefore, in CASP10, we participated with TSlab-psQA (pure-single Quality Assessment program) in QA category. Our goals were to estimate the quality score which has high correlation with the actual quality (GDT_TS) and to identify the best model from the model set, without using the consensus method. The TSlab-psQA estimates the quality of a single model based on the contact prediction method, secondary structure prediction method and the neural network training-prediction method without using any template information and consensus based score (i.e. pure-single model quality assessment method). Based on TSlab-psQA method, we developed another QA method TSlab-tbQA (quasi-single model method) that uses template-based evaluation score. Methods and Results of TSlab-tbQA are shown in the method abstract of TSlab-tbQA.

Methods

The psQA employed the four steps for each target as follows: (1) HHblits[1] was executed against UniProt20 to build a multiple sequence alignment (MSA). (2) The residue-residue contact prediction was performed from the MSA by PSICOV[2] , and the secondary structure prediction for each amino acid residues were performed by PSIPRED[3]. (3) From the features of local window (such as amino-acid type, secondary structures prediction, secondary structure and residue-residue contact prediction), an artificial neural network predicted the distance map of the all residues pairs of the target. In CASP10, we used the distance between the side-chain centers, not Cα atoms. The neural network was trained on 5664 targets (clustered PDB30 training data sets). (4) The estimated quality score was calculated by comparing the predicted distance map and the actual distance map of the model to be evaluated. All of the parameters of TSlab-psQA were optimized from CASP9 data and PDB30 data.

Results

Our preliminary results based on released 49/114targets in Sep 2012 were shown in the Table 1.

Table 1. Our preliminary analysis on 49 CASP10 targets

	Average(r)	**Overall(r)**	**Delta GDT_TS**	**#**
TSlab-psQA_stage1	0.62	0.47	7.9	44
TSlab-psQA_stage2	0.31	0.45	5.1	44

1. Remmert, M., Biegert, A., Hauser, A. & Soding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods 9, 173-5.
2. Jones, D. T., Buchan, D. W., Cozzetto, D. & Pontil, M. (2012). PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 28, 184-90.
3. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292, 195-202.

Conformation sampling with weakened repulsive force and selection from the ensemble

Hiromitsu Shimoyama, Genki Terashi , Yuuki Nakamura and Mayuko Takeda-Shitaka

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

We participated in the refinement category of CASP10. Our method can be divided into two parts. First, all-atom molecular dynamics (MD) simulations were performed in order to sample various conformations around starting structure. Second, qualities of the conformations were evaluated by using several descriptors. Here we describe original part of our method briefly.

Methods

Potential surfaces of proteins are expected to be rugged and complicated: for example, steric clashes among atoms cause such surfaces. In order to sweep conformation space quickly, our group performed MD with weakened repulsive force: Lenard-Jones potentials are partially linearized as follows.

$$U_{vdW}(r_{ij}) = \begin{cases} \varepsilon\left[(\sigma/r_{ij})^{12} - 2(\sigma/r_{ij})^{6}\right] \text{ (for } r_{ij} \geq r^{*}\text{: outer region)} \\ ar_{ij} + b \text{ (for } r_{ij} < r^{*}\text{ : inner region)} \end{cases}$$

ε is an energy coefficient, r_{ij} is a distance between atom i and j, σ is equilibrium distance. a is an energy coefficient and $b = -ar^{*} + U_{vdW}(r^{*})$. r^{*} is distance parameter at which outer- and inner-potential coincide with each other.

The energy coefficient a was taken to be a time-dependent parameter, i.e. $a(t) = a_0 + a_1\{1 + cos(\omega t)\}$. The parameters are taken to be $a_0 = 10$ kcal/mol/Å, $a_1 = 90$ kcal/mol/Å, and $\omega = 2\pi \times 10^2$ psec. Ordinary structures are probably obtained at least every $1/\omega$ seconds. In addition to the linearization, temperature was also increased every $1/\omega$ second, i.e. $T = T_0 + T_1\{1 - cos(\omega t)\}$: $T_0 = 300$ K and $T_1 = 400$ K. These methods were implemented in a program *myPresto* [1].

By these modifications, our MD simulation can sample more quickly than ordinary MD. Next problem is how to select better structures from the ensemble. In order to assess the quality, three parameters were considered, (1) TSlab-psQA score, (2) DFIRE[2] potential, (3) solvent accessible surface area (SAS). The TSlab-psQA score is a single-model quality assessment score which was developed and used for quality-assessment category by TSlab-psQA group (see detail in abstract by the group). SAS was used in order to select compact structure. Z-score of these variables was obtained from MD ensemble and averaged. We selected a structure of the highest averaged Z-score as the best structure.

References

1. Fukunishi,Y., Mikami,Y. & Nakamura,H (2003). The filling potential method: A method for estimating the free energy surface for protein-ligand docking. *J. Phys. Chem. B.* **107**, 13201-13210.

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

Single-model quality assessment method using template-based evaluation score

Yuuki Nakamura, Genki Terashi, Hiromitsu Shimoyama and Mayuko Takeda-Shitaka

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

In CASP10, we participated with TSlab-psQA and TSlab-tbQA in Quality Assessment (QA) category. TSlab-psQA estimates the quality of a single model based on the contact prediction method and the neural network training-prediction method without using any template information and consensus based score (pure-single model method) (see method abstract of TSlab-psQA). In order to improve accuracy to estimate the quality of the model that has high correlation with the actual quality, TSlab-tbQA (quasi-single model method) calculates the score by combining template-based evaluation score with the TSlab-psQA score.

Methods

TSlab-tbQA employed the three steps for each target as follows: (1) search for homologous templates: HHblits[1] was used to search sequence, alignment, and construct coarse structure models which were composed of only aligned Cα atoms. Each residue was weighted by confidence value of secondary structure estimated by PSIPRED[2]. (2) Calculate structural similarity score between the given model and homologous templates: the three coarse structure models from the top three templates were used as the homologous templates. As the structural similarity score, GDT[3] between the given model and the each coarse structure models was summed, and normalized to the dimension of the GDT_TS[3] value. (3) Calculate the final score by combining the above score with the TSlab-psQA score: the two scores were combined by linear combination. As dataset for the normalization and the linear combination, server models in the CASP9 were used.

Results

Our preliminary results (Table 1) show that TSlab-tbQA has better average correlation values than TSlab-psQA.

Table 1. Our preliminary analysis based on released 49/114 CASP10 targets in Sep 2012

	Average(r)	Overall(r)	Delta GDT_TS	#
TSlab-psQA_stage1	0.62	0.47	7.9	44
TSlab-psQA_stage2	0.31	0.45	5.1	44
TSlab-tbQA_stage1	0.70	0.84	5.7	49
TSlab-tbQA_stage2	0.40	0.82	6.8	49

1. Remmert, M., Biegert, A., Hauser, A. & Soding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* **9**, 173-5.
2. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* ***292***, 195-202.
3. Zemla, A. (2003). LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res.* ***31***, 3370–3374.

Multiplayer online game-based homology and ab-initio modeling

M. Gaebel[1], F. Khatib[2], J. Flatten[2], S. Cooper[2], T. Husain[2], K. Xu[2], Z. Popović[2], D. Baker[2] and Foldit Void Crushers Group[3]

[1] - *Admin and submitter of the Void Crushers,* [2] - *University of Washington, Seattle, WA,* [3] - *Worldwide*
gaebel@gmx.de

Models were constructed using Foldit, the online multiplayer game at http://fold.it. CASP10 targets shorter than 170 residues were given to Foldit players as puzzles to solve. Foldit allows players to form teams for cooperative gameplay; in this case predictions were selected from members of the Foldit team Void Crushers.

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 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 CASP10 puzzle was typically accessible to Foldit players for 8-9 days.

For CASP10 targets shorter than 170 residues in the "All Groups" category, two different Foldit puzzles were given to the players. One puzzle started from an extended chain, with alignments to known templates taken from the RAPTOR[3], SPARKS[4], and HHsearch[5] servers provided. 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 in-game tools listed above. For the second puzzle, models were constructed using the QUARK[6] and Zhang-Server[7] predictions. These server models were initially minimized using Rosetta and then given as starting points for the Foldit players to refine. This same protocol was used for CASP10 targets in the "Refinement" category, where server models were first minimized with Rosetta before being given to the Foldit players.

Quality and ranking of individual models was determined initially by the Rosetta full-atom energy. Submissions were then selected from Void Crushers predictions (that were not submitted from group FOLDIT already) based on:

- (for refinement targets) the fit between actual difference (GDT_TS) of the prediction from the starting model and expected difference of a good solution from the starting model. (GDT_TS calculated by TM-Score[8])
- the Rosetta energy score
- (for normal targets) the diversity from starting structures provided by servers and from other

models submitted by our group (compared visually and by RMSD through PyMOL[9])

- the probability of the secondary structures of a model (compared with the predictions by the SAM-T08 server[10])
- (for cysteine heavy targets) the number of disulfide bonds

For de-novo targets models folded from scratch or using templates did get boni compared to models based on server models.

Availability

Foldit is available through the Rosetta Commons at http://tinyurl.com/academic-foldit

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. *Methods in enzymology* **487**, 545-74.
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. Peng,J. & Xu,J. (2009) Boosting Protein Threading Accuracy. *Research in Computational Molecular Biology (RECOMB)*, **5541**, 31-45.
4. Yang,Y., Faraggi,E., Zhao, H. & Zhou,Y. (2011) Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of the query and corresponding native properties of templates. *Bioinformatics* **27**, 2076-2082.
5. Söding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics.* **21**(7), 951-60.
6. Xu,D. & Zhang,Y. (2012) Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins* **80**, 1715-35
7. Zhang,Y. (2008) I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*, vol **9**, 40
8. Zhang,Y., Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins.* **57,** 702-710.
9. The PyMOL Molecular Graphics System, Version 1.4.1 Schrödinger, LLC.
10. Karplus,K. (2009) SAM-T08: HMM-based Protein Structure Prediction. *Nucleic Acids Research.* **37**(2), W492-7

Protein Structure Prediction via Model Selection by APOLLO and Refinement by TASSER

Hongyi Zhou[1], Jilong Li[2], Xin Deng[2],Jianlin Cheng[2], Jeffrey Skolnick[1] and Silvia N. Crivelli[3].

[1] – *Center for the Study of Systems Biology, School of Biology, Georgia Institute of Technology*
[2] – *Department of Computer Science, University of Missouri-Columbia*
[3] – *Department of Computer Science, University of California, Davis*
hzhou3@gatech.edu

WeFold is a group branch of the WeFold collaboration (http://www.wefold.org). It tests the combination of model selection from all CASP servers by APOLLO model quality assessment prediction method and the TASSER refinement protocol.

Methods

All models from all CASP servers were assessed by the APOLLO model quality assessment prediction method. APOLLO[1] first filters out illegal characters and chain-break characters in the models predicted for a target. And then it performs a full pair-wise comparison between these models by calculating the GDT-TS scores between each pair of models using the tool TM-Score[2]. The average pair-wise GDT-TS score between a model and all other models is used as the predicted GDT-TS score of the model. Subsequently, TASSER[3] method was employed to refine the top 30 selected models. First, TASSER extracts distance and contact restraints based on consensus conformations of the 30 selected structures. Then, it starts from the 30 structures and moves them to satisfy the distance and contact restraints using replica exchange Monte Carlo simulation[4] and C_α representation. Low energy trajectories were output at fixed step intervenes. At the end of simulation, these trajectories were clustered using the SPICKER approach.[5] Models selected for submission are the top cluster centroids with rebuilt main-chain and side-chain atoms.

Availability

APOLLO (Quality-assessment) http://sysbio.rnet.missouri.edu/apollo/
TASSER (Refinement): http://cssb.biology.gatech.edu/
WeFold (Collaborative Protein Folding): http://www.wefold.org/

1. Wang, Z., Eickholt, J. & Cheng, J. (2011). APOLLO: A Quality Assessment Service for Single and Multiple Protein Models. *Bioinformatics* **27**, 1715-1716.
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. Zhang, Y. & Skolnick, J. (2004). Automated structure prediction of weakly homologous proteins on a genomic scale. Proceedings of the National Academy of Sciences of the United States of America 101, 7594-7599.
4. Gront, D., Kolinski, A. & Skolnick, J. (2001). A new combination of replica exchange Monte Carlo and histogram analysis for protein folding and thermodynamics. Journal of Computational Physics 115, 1569-1574.
5. Zhang, Y. & Skolnick, J. (2004). SPICKER: A clustering approach to identify near-native protein folds. Journal of Computational Chemistry 25, 865-871.

Application of Replica Exchange Molecular Dynamics with Implicit Solvation for Refinement of Collaboratively Generated and Ranked Models

L.O. Bortot[1], R.A. Faccioli[2], A.C.B. Delbem[2] & the WeFold Community[3]

1: *Laboratory of Biological Physics, Faculty of Pharmaceutical Sciences at Ribeirão Preto, University of São Paulo, Ribeirão Preto, São Paulo, Brazil. 2: Institute of Mathematic Science and Computer Science, University of São Paulo, São Carlos, São Paulo, Brazil. 3: http://www.wefold.org*

WeFold is an open collaboration initiative bringing together a dozen labs from all over the world through the science community gateway http://www.wefold.org. It enables the interaction among various groups that work on different components of the protein structure prediction pipeline, making it possible to leverage expertise at a scale that has never been done before. The collaboration resulted in five different branches, each submitting their own models. Here we describe the WeFoldMix branch.

The focus of this branch was to refine models which had already passed through all steps of the prediction pipeline as described in the other branches (wfFUIK, wfCPUNK, wfFUGT), e.g. prediction from hybrid homology/ab initio approaches, clusterization, refinement and ranking, each step using multiple methods and metrics. Specifically, we applied Replica Exchange Molecular Dynamics to the top ranked models for some targets aiming to further improve their quality.

Methods

A small set of high quality models collaboratively generated and ranked were chosen for applying the methodology described below. Each starting model was submitted to a two-step energy minimization with the steepest descent algorithm[1]. While in the first step no constraints were applied to the protein, in the second one all covalent bonds were constrained with the LINCS algorithm[2].

In conventional Molecular Dynamics (MD) simulations the atoms are moved along time according to the potential energy calculated with the equations and parameters of the chosen forcefield[3]. Because of the rugged nature of the potential energy landscape that describes the conformational behavior of proteins, they usually get trapped in local minima, which hinders the adequate sampling of the conformational space. The Replica Exchange Molecular Dynamics (REMD) algorithm tries to overcome this by allowing the system to diffuse through temperature space, facilitating the overcoming of potential energy barriers. When this algorithm is employed, multiple conventional MD simulations - called replicas - are done simultaneously at different temperatures and they are allowed to exchange temperatures at a given frequency according to Metropolis criterion[4-5].

We used a temperature range of 309 to 373K with 8 replicas. After 1 to 3 nanoseconds of REMD, the 309K trajectory portion which reached convergence was used for cluster analysis using a single linkage algorithm. Each cluster centroid was submitted to the same previously described two-step energy minimization process and each minimized cluster centroid was ranked based on several structural and energetic metrics, such as potential energy, number of intra-protein hydrogen bonds and hydrophobic solvent accessible surface area. Those with the best compromise among all the metrics were sent to CASP10.

All simulations were carried out with the GROMACS 4.5.5 simulation suite[6-7] using the AMBER99SB-ILDN forcefield[8]. In order to speed up the simulations the GBSA implicit solvation model[9] was used with the OBC algorithm for calculating the Born radii[10].

Availability
The GROMACS software suite is freely available at www.gromacs.org

1. van der Spoel,D., Lindahl,E., Hess,B., van Buuren,A.R., Apol,E., Meulenhoff,P.J., Tieleman,D.P., Sijbers,A.L.T.M., Feenstra,K.A., van Drunen,R. & Berendsen,H.J.C. (2010). GROMACS User Manual version 4.5.4, www.gromacs.org.
2. Hess,B., Bekker,H., Berendsen,H.J.C. & Fraaije,J.G.E.M. (1997). LINCS: A linear constraint solver for molecular simulations. *J. Comp. Chem*. **18**, 1463-1472.
3. Karplus,M. & Mccammon,A. (2002). Molecular Dynamics simulations of biomolecules. *Nature Structural Biology.* **8**, 646-652.
4. Sugita,Y. & Okamoto,Y. (1999). Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters.* **314(12)**, 141-151.
5. Lei,H. & Duan,Y. (2007). Improved sampling methods for molecular simulation. Current *Opinion in Structural Biology.* **17(2)**, 187-191.
6. Berendsen,H.J.C., Van Der Spoel,D. & Van Drunen,R. (1995). GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications.* **91(1-3)**, 43-56.
7. van der Spoel,D., Lindahl,E., Hess,B., Groenhof,G., Mark,A.E. & Berendsen,H.J.C. (2005). GROMACS: Fast, flexible and free. *J. Comp. Chem.* **26(16)**, 1701-1718.
8. Lindorff-Larsen,K., Piana,S., Palmo,K., Maragakis,P., Klepeis,J.L., Dror,R.O. & Shaw,D.E. (2010). Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins.* **78(8)**, 1950-1958.
9. Tsui,V. & Case,D.A., (2000). Theory and applications of the generalized born solvation model in macromolecular simulations. *Biopolymers.* **56(4)**, 275-291.
10. Onufriev,A., Bashford,D. & Case,D.A. (2004). Exploring protein native states and large-scale conformational changes with a modified Generalized Born model. *Proteins.* **55(2)**, 383-394.

Use of UNRES force field with secondary-structure and contact prediction in blind prediction of protein structure as part of the WeFold collaborative initiative

Yi He[1], Paweł Krupa[2], Adam K. Sieradzan[2], Magdalena Mozolewska[2], Tomasz Wirecki[2], George A. Khoury[3], Gaurav Chopra[4], James Smadbeck[3], Christodoulos A. Floudas[3], Michael Levitt[4], Harold A. Scheraga[1], Silvia Crivelli[5] and Adam Liwo[2*]

[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] – Department of Chemical and Biological Engineering, Princeton University, A325 Engineering Quadrangle, Princeton University, Princeton, NJ 08544, [4] – Department of Structural Biology, Stanford University School of Medicine, 299 Campus Drive West, D100 Fairchild Bldg., Stanford, CA 94305-5126, [5] – Department of Computer Science, UC Davis, 1 Shields Ave., Davis, CA 95616

adam@chem.univ.gda.pl

WeFold is an open collaboration initiative bringing together a dozen labs from all over the world through the science community gateway http://www.wefold.org. It enabled the interaction among groups that work on different components of the protein structure prediction pipeline, thus making it possible to leverage expertise at a scale that has not been done before. The collaboration resulted in five different branches, each submitting their own models. Here we describe the wfCPUNK branch.

Methods

The structures of the target proteins were predicted by a procedure which consists of the following four steps. First, coarse-grained simulations with the UNRES force field, with dihedral-angle and distance restraints imposed on the virtual-bond dihedral angles between the consecutive α-carbon (Cα) atoms and virtual side-chain distances, respectively, were employed to carry out Multiplexed Replica Exchange Molecular Dynamics (MREMD)[1] for target proteins. The restraints were obtained by using CONCORD[2] for secondary-structure prediction, BeST[3] for beta-sheet topology prediction, and a physics-based method of inter-residue contact prediction.[4; 5] Second, based on MREMD simulation results, Weighted-Histogram Analysis Method (WHAM) analysis was used to calculate relative free energy of each structure of last slice of MREMD simulation; the respective procedure is described in ref. 6. Third, cluster analysis was employed to cluster the structures from a MREMD simulation. Five clusters with lowest free energies were chosen as prediction candidates. The conformations closest to the respective average structures corresponding to the found clusters were converted to all-atom structures[7; 8] and energy minimized using a knowledge-based potential followed by stereochemical correction implemented in the KoBaMIN server.[9] This knowledge-based potential has been benchmarked using an extensive set of decoys[10] and previous CASP models.[11]

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 cluster-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. In

this CASP exercise we also introduced correlation terms that couple the backbone and side chain local-interaction energies. The effective energy function depends on temperature and has been parameterized to reproduce structure and thermodynamics of selected training proteins.[6; 12]

Availability

The components of the methods used are available by the developers of each method at the following webpages.

UNRES (Package to perform coarse-grained simulations of protein structure and dynamics): http://www.unres.pl

CONCORD (2°structure prediction): http://helios.princeton.edu/CONCORD/

BeST (β-sheet topology prediction): http://selene.princeton.edu/BeST/

KoBaMIN (refinement): http://csb.stanford.edu/kobamin

WeFold (Collaborative protein folding effort): http://www.wefold.org

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 α and α+β Proteins. *Journal of Chemical Theory and Computation* **5**, 627-640.
2. Wei, Y., Thompson, J. & Floudas, C. A. (2012). CONCORD: a consensus method for protein secondary structure prediction via mixed integer linear optimization. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science* **468**, 831-850.
3. Subramani, A. & Floudas, C. A. (2012). β-sheet Topology Prediction with High Precision and Recall for β and Mixed α/β Proteins. *PLoS ONE* **7**, e32461.
4. Rajgaria, R., Wei, Y. & 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**, 1825-46.
5. Rajgaria, R., McAllister, S. R. & Floudas, C. A. (2009). Towards accurate residue–residue hydrophobic contact prediction for α helical proteins via integer linear optimization. *Proteins: Structure, Function, and Bioinformatics* **74**, 929-947.
6. Liwo, A., Khalili, M., Czaplewski, C., Kalinowski, S., Ołdziej, S., Wachucik, K. & Scheraga, H. A. (2006). 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. *The Journal of Physical Chemistry B* **111**, 260-285.
7. Kazmierkiewicz, R., Liwo, A. & Scheraga, H. A. (2003). Addition of side chains to a known backbone with defined side-chain centroids. *Biophysical Chemistry* **100**, 261-280.
8. Kazmierkiewicz, R., Liwo, A. & Scheraga, H. A. (2002). Energy-based reconstruction of a protein backbone from its α-carbon trace by a Monte-Carlo method. *Journal of Computational Chemistry* **23**, 715-723.
9. Rodrigues, J. P. G. L. M., Levitt, M. & Chopra, G. (2012). KoBaMIN: a knowledge-based minimization web server for protein structure refinement. *Nucleic Acids Research* **40**, W323-W328.
10. Chopra, G., Summa, C. M. & Levitt, M. (2008). Solvent dramatically affects protein structure refinement. *Proceedings of the National Academy of Sciences of the USA*, **105**, 20239-20244.

11. Chopra, G., Kalisman, N. & Levitt, M. (2010). Consistent refinement of submitted models at CASP using a knowledge-based potential. *Proteins: Structure, Function, and Bioinformatics* **78**, 2668-2678.
12. 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. *Journal of Computational Chemistry* **30**, 2127-2135.

Hybrid Human Protein Structure Prediction via the Online Multiplayer Game Foldit Coupled with an Iterative Clustering Approach for Selection of Near-Native Structures, Knowledge-Based Refinement, and State-of-the-Art Scoring Functions

George A. Khoury[1], Firas Khatib[2], Hongyi Zhou[3], Gaurav Chopra[4], Jilong Li[5], Seth Cooper[6], Jeff Flatten[6], Tamir Husain[6], Kefan Xu[6], James Smadbeck[1], Xin Deng[5], Zoran Popović[6], Jianlin Cheng[5], Michael Levitt[4], Jeffrey Skolnick[3], David Baker[2], Christodoulos A. Floudas[1], Silvia N. Crivelli[7], and Foldit Players[8]

[1] – *Department of Chemical and Biological Engineering, Princeton University*
[2] – *Department of Biochemistry, University of Washington*
[3] – *Center for the Study of Systems Biology, School of Biology, Georgia Institute of Technology*
[4] – *Department of Structural Biology, Stanford University School of Medicine*
[5] – *Department of Computer Science, University of Missouri-Columbia*
[6] – *Center for Game Science, Department of Computer Science & Engineering, University of Washington*
[7] – *Department of Computer Science, University of California, Davis*
[8] – *Worldwide*
george@titan.princeton.edu

A hybrid method using structures constructed by human players via Foldit and selected by a consensus of state-of-the-art in-house computational methods from 7 research groups was used for the prediction of human and refinement CASP targets as part of the WeFold collaboration. WeFold is an open collaboration initiative bringing together a dozen labs from all over the world through the science community gateway http://www.wefold.org. It enables the interaction between groups that work on different components of the protein structure prediction pipeline, making it possible to leverage expertise at a scale that has not been done before. The collaboration resulted in five different branches, each submitting their own models. Here we describe the wfFUIK branch.

Team members from the different labs adapted the in-house methods designed to operate on smaller datasets to be able to tractably perform the corresponding calculations with the large datasets within each deadline. Furthermore, methods were adapted to be able to handle systems containing structural symmetry. The methodology for filtering, clustering, ranking, and selection was conceived collaboratively using the strengths of each contributing group.

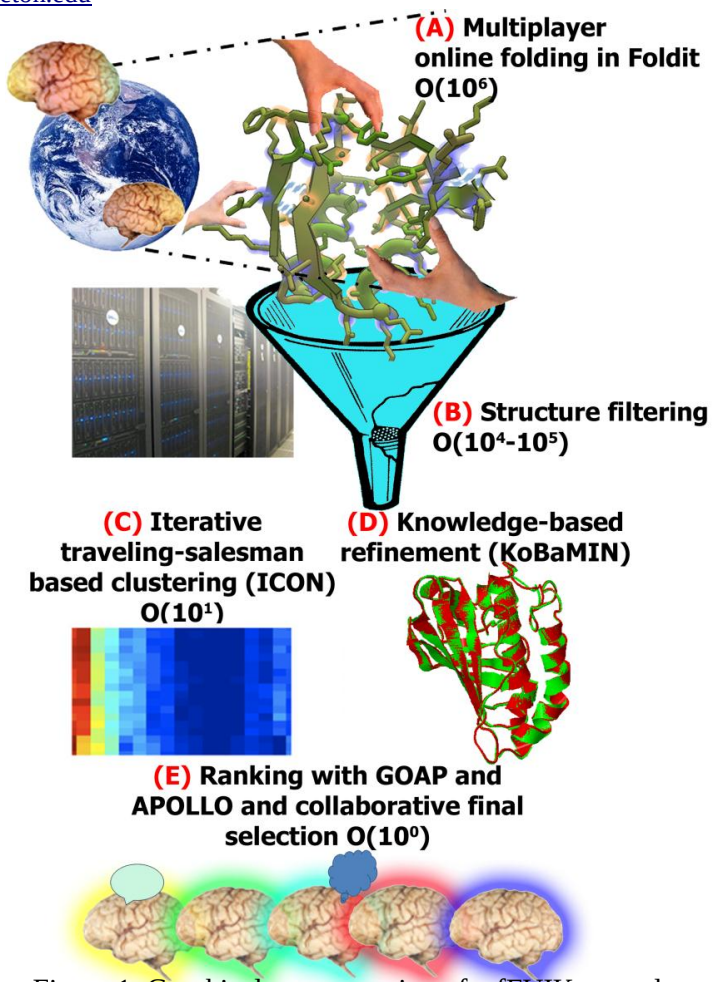


Figure 1: Graphical representation of wfFUIK procedure

Methods

Figure 1 represents the combined methodology. (A) Human players generated an ensemble of protein models using the online multiplayer game Foldit[1] (http://fold.it) on the order of 10^6 models per target. (B) Structural filtering was performed to eliminate duplicates (RMSD ≤ cutoff) as well as those with unrealistic SASAs, and those lacking relevant secondary structure elements, resulting in an enriched set on the order of 10^4-10^5 structures. (C) The iterative traveling salesman based clustering algorithm, ICON[2] was used to select less than 100 models representing the entire conformational space including the lowest energy structures based on the Rosetta[3] and dDFIRE[4] energy functions. (D) These models were refined using a knowledge-based potential followed by stereochemical correction implemented in the KoBaMIN[5; 6; 7] server. (E) Finally, GOAP,[8] an orientation-dependent, all-atom statistical potential and APOLLO,[9] a quality-assessment method were used to rank the models, leading to a consensus.

Availability

The individual methods contributing to the collaborative effort (http://www.wefold.org) are available online:
Foldit (Online multi-player game to solve folding puzzles): http://fold.it
Rosetta (Scientific machinery behind Foldit): http://www.rosettacommons.org/
ICON (TSP-based clustering): http://helios.princeton.edu/ICON/
KoBaMIN (Refinement): http://csb.stanford.edu/kobamin
GOAP (Scoring): http://cssb.biology.gatech.edu/GOAP/index.html
APOLLO (Quality-assessment): http://sysbio.rnet.missouri.edu/apollo/

1. 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.
2. Subramani, A., DiMaggio, P. A. & Floudas, C. A. (2009). Selecting High Quality Protein Structures from Diverse Conformational Ensembles. *Biophysical Journal* **97**, 1728-1736.
3. Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P. D., Smith, C. A., Sheffler, W., Davis, I. W., Cooper, S., Treuille, A., Mandell, D. J., Richter, F., Ban, Y. E., Fleishman, S. J., Corn, J. E., Kim, D. E., Lyskov, S., Berrondo, M., Mentzer, S., Popovic, Z., Havranek, J. J., Karanicolas, J., Das, R., Meiler, J., Kortemme, T., Gray, J. J., Kuhlman, B., Baker, D. & Bradley, P. (2011). ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods in enzymology* **487**, 545-74.
4. Yang, Y. & Zhou, Y. (2008). Specific interactions for ab initio folding of protein terminal regions with secondary structures. *Proteins: Structure, Function, and Bioinformatics* **72**, 793-803.
5. Rodrigues, J. P. G. L. M., Levitt, M. & Chopra, G. (2012). KoBaMIN: a knowledge-based minimization web server for protein structure refinement. *Nucleic Acids Research* **40**, W323-W328.
6. Chopra, G., Summa, C. M. & Levitt, M. (2008). Solvent dramatically affects protein structure refinement. *Proceedings of the National Academy of Sciences*.
7. Chopra, G., Kalisman, N. & Levitt, M. (2010). Consistent refinement of submitted models at CASP using a knowledge-based potential. *Proteins: Structure, Function, and Bioinformatics* **78**, 2668-2678.

8. Zhou, H. & Skolnick, J. (2011). GOAP: A Generalized Orientation-Dependent, All-Atom Statistical Potential for Protein Structure Prediction. *Biophysical Journal* **101**, 2043-2052.
9. Wang, Z., Eickholt, J. & Cheng, J. (2011). APOLLO: A Quality Assessment Service for Single and Multiple Protein Models. *Bioinformatics* **27**, 1715-1716.

Foldit with Selection by Knowledge-based Potential GOAP and Refinement by TASSER

Hongyi Zhou[1], Firas Khatib[2], George A. Khoury[3], Seth Cooper[4], Jeff Flatten[4], Tamir Husain[4], Kefan Xu[4], James Smadbeck[3], Zoran Popović[4], Christodoulos A. Floudas[3], David Baker[2], Jeffrey Skolnick[1], and Foldit Players[5]

[1] – *Center for the Study of Systems Biology, School of Biology, Georgia Institute of Technology*
[2] – *Department of Biochemistry, University of Washington*
[3] – *Department of Chemical and Biological Engineering, Princeton University*
[4] – *Center for Game Science, Department of Computer Science & Engineering, University of Washington*
[5] – *Worldwide*
hzhou3@gatech.edu

This branch of the WeFold collaborative protein folding initiative (http://www.wefold.org) is similar to the branch wfFUIK in that it starts with a filtered set of protein structures generated by the online multiplayer game Foldit. However, it deviates from that branch in the subsequent steps to select and refine the chosen models. The wfFUGT branch tests the combination of sampling by human Foldit players coupled with filtering algorithms, model selection by the knowledge-based potential GOAP (not of a consensus-based kind such as clustering) and the TASSER refinement protocol. These methods are from different active groups and the experiment would be impossible without the Wefold collaboration.

Methods

Humans from all over the world playing the online multiplayer game Foldit[1] first generate protein structure models (http://fold.it). This step produces models on the order of 10^6 per target. A structural filtering step was then performed to eliminate duplicate structures (RMSD $\leq$ cutoff), structures with unrealistic solvent-accessible surface areas,[2] and structures lacking relevant or any secondary structure elements.[3] This resulted in an enriched set of models on the order of 10^4-10^5 structures. Subsequently, the knowledge-based potential GOAP[4] was used to select the top 30 models from the enriched set. TASSER[5] was employed to refine the selected models. TASSER is primarily developed for refining template models built upon PDB structures found by threading methods. Here, we applied it to artificially generated Foldit structures driven by the knowledge-based potential Rosetta.[6] First, it extracts distance and contact restraints based on consensus conformations of the 30 selected structures. Then, it starts from the 30 structures and moves them to satisfy the distance and contact restraints using replica exchange Monte Carlo simulation[7] and C_α representation. Low energy trajectories were output at fixed step intervenes. At the end of simulation, these trajectories were clustered using the SPICKER approach.[8] Models selected for submission are the top cluster centroids with rebuilt main-chain and side-chain atoms.

Availability

The individual methods contributing to the wfFUGT collaborative effort are available online at the following websites:
Foldit (Online multi-player game to solve folding puzzles): http://fold.it/
Rosetta (Scientific machinery behind Foldit): http://www.rosettacommons.org/

GOAP (Selection): http://cssb.biology.gatech.edu/GOAP/index.html
TASSER (Refinement): http://cssb.biology.gatech.edu/
WeFold (Collaborative Protein Folding): http://www.wefold.org/

1. 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.
2. Hubbard, S. J. & Thornton, J. M. (1993). 'NACCESS', computer program.
3. Frishman, D. & Argos, P. (1995). Knowledge-based protein secondary structure assignment. Proteins: Structure, Function, and Bioinformatics 23, 566-579.
4. Zhou, H. & Skolnick, J. (2011). GOAP: A Generalized Orientation-Dependent, All-Atom Statistical Potential for Protein Structure Prediction. Biophysical Journal 101, 2043-2052.
5. Zhang, Y. & Skolnick, J. (2004). Automated structure prediction of weakly homologous proteins on a genomic scale. Proceedings of the National Academy of Sciences of the United States of America 101, 7594-7599.
6. Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P. D., Smith, C. A., Sheffler, W., Davis, I. W., Cooper, S., Treuille, A., Mandell, D. J., Richter, F., Ban, Y. E., Fleishman, S. J., Corn, J. E., Kim, D. E., Lyskov, S., Berrondo, M., Mentzer, S., Popovic, Z., Havranek, J. J., Karanicolas, J., Das, R., Meiler, J., Kortemme, T., Gray, J. J., Kuhlman, B., Baker, D. & Bradley, P. (2011). ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods in enzymology 487, 545-74.
7. Gront, D., Kolinski, A. & Skolnick, J. (2001). A new combination of replica exchange Monte Carlo and histogram analysis for protein folding and thermodynamics. Journal of Computational Physics 115, 1569-1574.
8. Zhang, Y. & Skolnick, J. (2004). SPICKER: A clustering approach to identify near-native protein folds. Journal of Computational Chemistry 25, 865-871.

The YASARA homology modeling module V3.0 with new 'PSSP alignments' and model hybridization from multiple templates

E. Krieger and G.Vriend

CMBI 260, NCMLS, Radboud University Nijmegen Medical Center,
PO Box 9101, 6500 HB Nijmegen, the Netherlands,
elmar@cmbi.ru.nl

Like in CASP9, the 'YASARA Structure' server (*www.yasara.org/homologymodeling*) submitted predictions for those targets that could be built reliably using known template structures. CASP9 evaluation identified alignment accuracy and model hybridization from multiple templates as the main bottlenecks, which have therefore been the development focus, while hires refinement needed less attention[1].

Methods

As in previous CASPs, our method targets classic homology modeling with a focus on high-resolution refinement. This involves running PsiBLAST with Uniref90 profiles to identify the top 20 templates, using stochastic[2] profile-profile alignments including SSALN features[3] to arrive at alternative high-scoring target-template alignments, building models for all of them (using SCWRL[4] rotamer libraries, but additional energy terms), scoring them, and fusing the best parts to a hybrid model. For CASP10, the focus was on better template profiles, which are now based on 'PSSP files' (**P**rofiles from **S**equence- and **S**tructurally related **P**roteins), that use twisted structural alignments to go beyond what HSSP files offer (soon available for download). Additionally, a new hybrid modeler was developed, which combines the best parts from multiple models to hopefully get closer to the target. 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 homology models with reliable quality scores for 77 CASP10 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 including 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**

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. Mueckstein U, Hofacker IL and Stadler PF (2002). Stochastic pairwise alignments. *Bioinformatics* 18 Sup2, 153-160

3. 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
4. Canutescu AA, Shelenkov AA and Dunbrack RL Jr. (2003). A graph-theory algorithm for rapid protein side-chain prediction.
Protein Sci. **12**, 2001-2014.

Protein structure predictions by a combination of I-TASSER and QUARK pipelines

Yang Zhang, Dong Xu, Jianyi Yang, Ambrish Roy, Renxiang Yan

Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA
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[1; 2; 3] which are essentially the same, except for that the human prediction exploited the templates in CASP10 Server Section, while Zhang-Server used our in-house threading programs. QUARK is a pipeline developed for *ab initio* protein folding,[4; 5] which was recently extended to template-assisted structure assembly (Xu & Zhang, in preparation). All the procedures are fully automated, in the sense that no human intervention is needed.

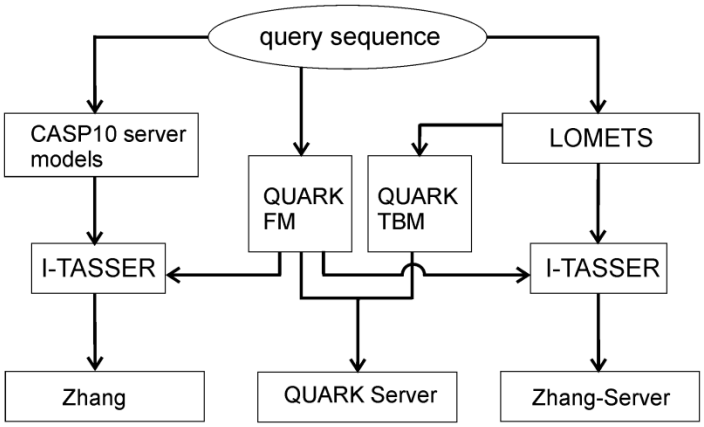


Figure 1. Flowchart for automated structure modeling generated for "Zhang", "Zhang-Server", and "QUARK" in CASP10.

Compared to our previous prediction procedures,[2; 6; 7] the major new developments in CASP10 are in the *ab initio* modeling procedure where we found that combining QUARK with threading alignments can improve model quality for both free-modeling (FM) and template-based modeling (TBM) targets. In the model selection, a new MQAP procedure is developed for final model selection, which includes multiple consensus- and physics-based model selectors.

The overall structure prediction pipelines include three general steps: template identification, structure re-assembly, 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,[8] a meta-server approach containing 8 locally installed threading programs. In human prediction, we additionally include 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 and *ab initio* structure assembly. The template-based modeling is mainly implemented by I-TASSER, where continuous fragments excised from the threading templates are exploited to assemble full-length models[1; 3; 9] with unaligned loop regions built by *ab initio* modeling.[10] The simulations are implemented in a modified replica-exchange Monte Carlo protocol.[11] The I-TASSER potential includes four components: (1) general knowledge-based statistics terms from the PDB (Cα/side-chain correlations[10], H-bond[12] and hydrophobicity[13]); (2) spatial restraints from threading templates[8]; (3) sequence-based Cα contact predictions by SVMSEQ;[14; 15] (4) distance map from segmental threading[16].

QUARK[4; 5] was originally developed for *ab initio* protein structure prediction without using global template structures, where short fragments of 1-20 residues are taken from unrelated proteins which are used to assemble the structural models under the guide of an optimized knowledge-based force field containing general statistical potentials and a protein-specific distance profile potential extracted from short fragments. In the new development, spatial restraints extracted from the LOMETS threading alignments are exploited to assist the QUARK structural assembly simulations. Depending on how the templates and restraints are used, four different version of QUARK programs were implemented in CASP10, i.e. QUARK-I: the default simulation without using threading templates; QUARK-II: default simulations but with initial conformation starting from threading template; QUARK-III: similar to QUARK-II but with distance profile restraints taken from the threading alignments; QUARK-IV: similar to QUARK-III but with the full-set of spatial restraints (Cα distance map and side-chain contacts, similar to I-TASSER restraints[3]) exploited in QUARK simulations.

Different procedures were used to generate models for different category of protein targets. In the QUARK server, for Hard targets, the programs QUARK-I and II are implemented; for Easy targets, QUARK-III and IV are implemented. In Zhang and Zhang-Server, for Hard targets, the models generated by QUARK-I and II simulations are used to sort the LOMETS templates, where the top templates which are structurally closest to the QUARK *ab initio* models are used by I-TASSER for the further structure assembly; for Easy targets, the default I-TASSER simulations are implemented to generate the structural decoys with the QUARK TBM models added in the starting conformation pool which are treated as 9th set of threading templates in addition to LOMETS templates (see Figure 1).

Model selection and refinements. The structures in low-temperature replicas of I-TASSER and QUARK simulations are clustered by SPICKER.[17] The atomic models are constructed by REMO[18] 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[19] and ModRefiner[20] for structure refinement, with the purpose of improving local geometry and H-bonding, and reducing steric clashes of the models.

To select models generated from different pipelines, we implement a set of seven MQAP programs, including the I-TASSER C-score, structural consensus measured by pair-wise TM-score, and five statistical potentials (RW, RWplus, Dfire, Dope and verify3D). Finally, a MQAP consensus score is defined as the sum of the rank of the seven MQAP scores and models of the lowest consensus scores are finally selected for submission.

Availability

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. Wu, S., Skolnick, J. & Zhang, Y. (2007). Ab initio modeling of small proteins by iterative TASSER simulations. *BMC Biol* **5**, 17.
2. Zhang, Y. (2007). Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins* **69**, 108-117.
3. Roy, A., Kucukural, A. & Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc* **5**, 725-38.
4. Xu, D. & Zhang, Y. (2012). Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins* **80**, 1715-35.
5. Xu, D. & Zhang, Y. (2012). Towards optimal fragment generations for ab initio protein structure assembly. *Proteins* (in press).
6. Zhang, Y. (2009). I-TASSER: Fully automated protein structure prediction in CASP8. *Proteins* **77**, 100-113.
7. Xu, D., Zhang, J., Roy, A. & Zhang, Y. (2011). Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based ab initio folding and FG-MD-based structure refinement. *Proteins* **79 (Suppl 10)**, 147-160.
8. Wu, S. T. & Zhang, Y. (2007). LOMETS: A local meta-threading-server for protein structure prediction. *Nucleic Acids Res.* **35**, 3375-3382.
9. Zhang, Y. & Skolnick, J. (2004). Automated structure prediction of weakly homologous proteins on a genomic scale. *Proc. Natl. Acad. Sci. USA* **101**, 7594-7599.
10. Zhang, Y., Kolinski, A. & Skolnick, J. (2003). TOUCHSTONE II: A new approach to ab initio protein structure prediction. *Biophys. J.* **85**, 1145-1164.
11. Zhang, Y., Kihara, D. & Skolnick, J. (2002). Local energy landscape flattening: Parallel hyperbolic Monte Carlo sampling of protein folding. *Proteins* **48**, 192-201.
12. Zhang, Y., Hubner, I., Arakaki, A., Shakhnovich, E. & Skolnick, J. (2006). On the origin and completeness of highly likely single domain protein structures *Proc. Natl. Acad. Sci. USA* **103**, 2605-10.
13. Chen, H. & Zhou, H. X. (2005). Prediction of solvent accessibility and sites of deleterious mutations from protein sequence. *Nucleic Acids Res* **33**, 3193-9.
14. Wu, S. & Zhang, Y. (2008). A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. *Bioinformatics* **24**, 924-31.
15. Wu, S., Szilagyi, A. & Zhang, Y. (2011). Improving protein structure prediction using multiple sequence-based contact predictions. *Structure* **19**, 1182-91.
16. Wu, S. & Zhang, Y. (2010). Recognizing protein substructure similarity using segmental threading. *Structure* **18**, 858-67.
17. Zhang, Y. & Skolnick, J. (2004). SPICKER: A clustering approach to identify near-native protein folds. *J Comput Chem* **25**, 865-71.
18. 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.
19. Zhang, J. & Zhang, Y. (2010). High-resolution protein structure refinement using fragment guided molecular dynamics. Submitted.
20. Xu, D. & Zhang, Y. (2011). Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. *Biophys J* **101**, 2525-34.

COACH: a consensus-based approach for protein ligand binding sites prediction

Jianyi Yang, Ambrish Roy and Yang Zhang

Department of Computational Medicine and Bioinformatics,
University of Michigan, 100 Washtenaw Avenue, Ann Arbor, MI 48109-2218, USA
zhng@umich.edu

The binding sites predictions for our human ("Zhang_FUNCTION") and server ("I-TASSER_FUNCTION") predictions are both based on COACH, a consensus approach for ligand-binding sites prediction. Since some components of COACH relies on 3D structure of the target proteins, the Zhang_FUNCTION used multiple 3D models predicted by other servers in CASP10 Server Section and the I-TASSER_FUNCTION exploited only the 3D models predicted by the "Zhang-Server".

COACH combines the binding-site prediction results of five methods: COFACTOR[1,2], FINDSITE[3], ConCavity[4], TMSITE, and SSITE. The first three are published methods and have been systematically benchmarked in a recent study[1]. TMSITE and SSITE are two recently developed methods to predict ligand-binding sites by the complementary structural alignment and sequence profile-profile alignment search, respectively. A recently developed database BioLiP[5] for biologically relevant ligand-protein interaction is used as the template library for COFACTOR, TMSITE and SSITE. The overall architecture of COACH is demonstrated in Figure 1.

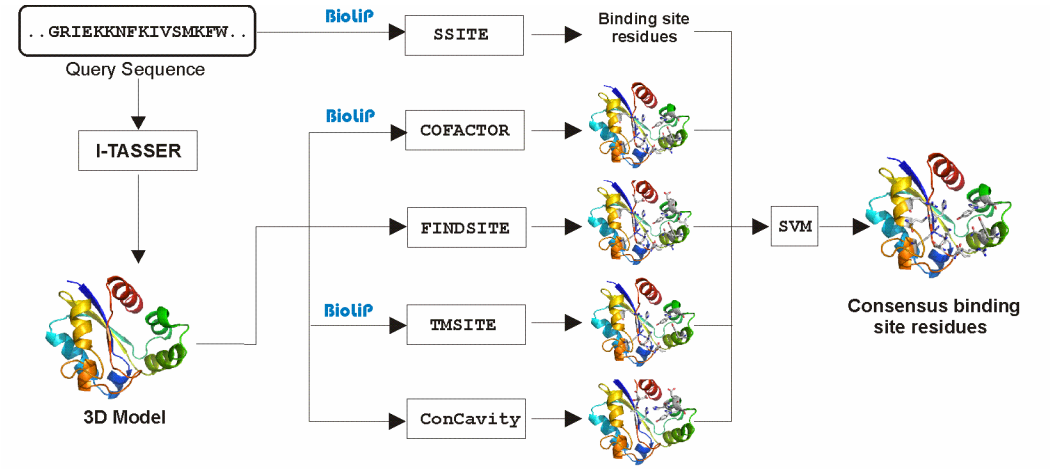


Figure 1. The architecture of COACH for binding sites prediction.

Given a query sequence, the structure prediction pipeline I-TASSER[6] is used to build 3D structure model. At the same time, the query sequence is submitted to the sequence-based approach SSITE to predict its binding site residues. Once the structure modeling is done, query's 3D models are submitted to four structure-based methods to predict binding site residues. The top predictions from COFACTOR, TMSITE, SSITE, FINDISTE, and ConCavity are then combined using Support Vector Machine (SVM). The probability of each residue to be a binding

site residue provided by each method is collected, which is used to construct a feature vector for each residue. Finally, the feature vector is fed into SVM to make consensus prediction.

Availability

The COACH algorithm is available at http://zhanglab.ccmb.med.umich.edu/BioLiP/coach.html

1. A. Roy, J. Yang, and Y. Zhang. COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Res.*, 40:W471–W477, 2012.
2. A. Roy and Y. Zhang. Recognizing protein-ligand binding sites by global structural alignment and local geometry refinement. *Structure*, 20:987–997, 2012.
3. M. Brylinski and J. Skolnick. A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proc. Natl Acad. Sci. USA*, 105:129–134, 2008.
4. J.A. Capra, R.A. Laskowski, J.M. Thornton, M. Singh, and T.A. Funkhouser. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Comput. Biol.*, 5:e1000585, 2009.
5. J. Yang, A. Roy, and Y. Zhang. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Submitted*, 2012.
6. A. Roy, A. Kucukural, and Y. Zhang, I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.*, 5:725-738, 2012.
7. J.A. Capra and M. Singh. Predicting functionally important residues from sequence conservation. *Bioinformatics*, 23:1875–1882, 2007.

CASP10 predictions using EdaFold

David Simoncini, Arnout R.D. Voet, Kam Y. J. Zhang

Zhang Initiative Research Unit, Advanced Science Institute, RIKEN, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan
simoncini@riken.jp, arnout.voet@riken.jp, kamzhang@riken.jp

Conformational sampling is one of the bottlenecks in fragment-based protein structure prediction approaches. They generally start with a coarse-grained optimization where main chain atoms and centroids of side chains are considered, followed by a fine-grained optimization with an all-atom representation of proteins. It is during this coarse-grained phase that fragment-based methods sample intensely the conformational space. If the native-like region is sampled more, the accuracy of the final all-atom predictions may be improved accordingly. We have proposed a new method for fragment-based protein structure prediction based on an Estimation of Distribution Algorithm, which we refer to as EdaFold[1].

Methods

EdaFold is a fragment-based protein structure prediction algorithm. Similarly to Rosetta[2], it is decomposed in two stages. First, 9-mers followed by 3-mers are assembled together to create coarse-grained models. 9-mers and 3-mers are taken from a fragment library which is created from protein structures available in the PDB. The fragment library we used was constructed using Rosetta's fragment picking method[3]. During the second stage, models are represented at atomic detail, and side chains are packed to minimize an all atom energy function. We use Rosetta Relax protocol to perform this operation.

EdaFold is an iterative process. Instead of randomly selecting fragments from the library, we use the Estimation of Distribution Algorithm to learn from previously generated decoys and steer the search toward native-like regions by constructing a non-uniform probability mass functions over the fragment library. At each iteration, the probabilities of selecting fragments for insertion are updated according to the observed frequency of each fragment in low energy models from the previous iteration. We perform 4 iterations, and models generated at each of them are present in the final population. At the first iteration, the fragments are selected randomly with a uniform probability mass function. This probability mass function is then updated according to the frequency of occurrences of each fragment in a subset of 15% lowest energy all-atom models. At the next iteration, fragments will be picked according to this new probability mass function and used to generate coarse-grained models. The coarse-grained models will be refined by an energy minimization process that relies on simulated annealing and iterated hill climbing. All the coarse-grained models will be turned into all-atom models and refined by the fast relax protocol in Rosetta.

EdaFold was used to predict the structures of targets from the “all groups” category in CASP10. We submitted models for 46 targets of this category. The number of models generated for each target ranges from 35 000 to 200 000 depending on the length of the sequence and computational resources. The models were clustered with a 3 Angstrom radius. One model was selected out of each of the top 5 clusters after visual inspection. When no cluster could be identified, models were ranked by energy. Up to 5 models were selected out of the 200 lowest energies after visual inspection.

Availability

The source code of a version of EdaFold producing coarse-grained models is available on our website: http://www.riken.jp/zhangiru/software.html

1. Simoncini, D., Berenger, F., Shrestha, R., and Zhang, K. Y. J. (2012). A probabilistic fragment-based protein structure prediction algorithm. PLoS One, 7(7), e38799.
2. Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P. D., Smith, C. A., Sheffler, W., Davis, I. W., Cooper, S., Treuille, A., Mandell, D. J., Richter, F., Ban, Y.-E. A., Fleishman, S. J., Corn, J. E., Kim, D. E., Lyskov, S., Berrondo, M., Mentzer, S., Popović, Z., Havranek, J. J., Karani-colas, J., Das, R., Meiler, J., Kortemme, T., Gray, J. J., Kuhlman, B., Baker, D., and Bradley, P. (2011). Rosetta3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol, 487, 545–74.
3. Gront, D., Kulp, D. W., Vernon, R. M., Strauss, C. E. M., and Baker, D. (2011). Generalized fragment picking in rosetta: design, protocols and applications. PloS One, 6(8), e23294.

Ab initio protein structure prediction using QUARK as guided by distance and contact restraints

D. Xu and Y. Zhang

Department of Computational Medicine and Bioinformatics, University of Michigan, USA
xudong@umich.edu and zhng@umich.edu

QUARK program was originally developed for template-free protein structure prediction.[1] We have recently extended it for template-based modeling as well. In CASP10, we keep using the default program for modeling targets which are defined as "hard" by LOMETS.[2] For other targets, QUARK starts from threading templates and use external distance and contact restraints as an energy term for restricted Monte Carlo structural assembly simulations. In the following, we mainly focus on the description of the second procedure.

Methods

Given the multiple threading templates by LOMETS, which have often alignment gaps, we first build the full-length backbone models by filling the gaps using a random walk procedure. QUARK then treats each of full-length models as the initial conformation of individual replicas in the Replica-exchange Monte Carlo simulation. Alternatively, we use up to 150 server models in Stage 2 as initial models of QUARK, when the CASP server models become available.

For the targets with threading templates, the distance profiles are extracted from multiple threading alignments; this term was obtained from the *ab initio* short fragments in the default QUARK program.[3] An energy term is designed in QUARK for evaluating the fitness of the decoy structures with the distance profiles.

If native distance/contact restraint information is provided (i.e. in the Contact-Assisted target category), we will add the information as an additional energy term to guide the QUARK simulation. During the procedure, we also manually check the distance profiles to examine whether some residue pairs have high probability to form beta pairs. The distance restraint data are then extracted from the distance profiles of the residue pairs which are used for the restricted Monte Carlo simulation.

Free-modeling targets are the major focus of this human group. We run four independent QUARK simulations based on the availability of distance/contact restraints.

(1) Default QUARK free modeling

(2) QUARK + distance restraints

(3) QUARK + LOMETS templates as initial conformations + distance restraints

(4) QUARK + CASP server models as initial conformations + distance restraints

After the Monte Carlo simulation, we use SPICKER[4] to generate five cluster centers by clustering all structural decoys. Since these models contain only backbone atoms, ModRefiner[5] is used to build the full-atomic models and refine their physical quality.

Results

We examined the modeling results for the free-modeling targets that have experimental structure released by the time this abstract was prepared. With the help of a few true contact restraints

which were provided for the Contact-Assisted targets, several free-modeling targets converted from 'non-foldable' to 'foldable' in our procedure (see Table 1).

Table 1. TM-score of the best QUARK model with and without true contact restraints.

Target	Without	With	Target	Without	With
Tc649	0.32	0.43	Tc658-D1	0.24	0.49
Tc673	0.39	0.48	Tc676	0.27	0.43
Tc678	0.36	0.64	Tc680	0.70	0.84
Tc705-D1	0.58	0.61	Tc735-D1	0.26	0.56
Tc735-D2	0.37	0.44			

For Tc680 which is a tetramer protein, we first modeled the monomer structure as guided by default intra-chain distance restraints. In the second step, a linker of 34 alanines was generated to connect each of the monomer pairs so that the tetramer could be treated as an artificial monomer in our simulation. During the simulations, the inter-chain distance restraints were converted to intra-chain distance restraints. Each of the monomers was kept rigid and only the linker region was flexible. The best tetramer model for this target has a TM-score=0.65 to the native.

Availability

The default QUARK prediction and that with by distance/contact restraints are available as an online server at http://zhanglab.ccmb.med.umich.edu/QUARK/.

1. Xu D, Zhang Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins 2012;80(7):1715-1735.
2. Wu S, Zhang Y. LOMETS: a local meta-threading-server for protein structure prediction. Nucleic Acids Res 2007;35(10):3375-3382.
3. Xu D, Zhang Y. Towards optimal fragment generations for ab initio protein structure assembly. Proteins 2012.
4. Zhang Y, Skolnick J. SPICKER: a clustering approach to identify near-native protein folds. J Comput Chem 2004;25(6):865-871.
5. Xu D, Zhang Y. Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. Biophys J 2011;101(10):2525-2534.

Hybrid structural refinement using ModRefiner and FG-MD

D. Xu and Y. Zhang

Department of Computational Medicine and Bioinformatics, University of Michigan, MI, USA
xudong@umich.edu and zhng@umich.edu

This group ('Zhang_Refinement') mainly focuses on the refinement category of the CASP10 although 3D models were submitted for all targets. Two programs, ModRefiner[1] and FG-MD[2], were implemented to refine the protein structures.

ModRefiner is a hierarchical program for protein structure refinement. It first constructs and refines the backbone structures with the consideration of only backbone atoms and side-chain center. In the second step, the full-atomic structures are relaxed during the repacking simulations (named 'Flexible ModRefiner'). It has an option to allow input of a reference model, where distance/contact restraints can be extracted, to guide the refinement simulations (named 'Restrained ModRefiner').

FG-MD refines the models based on molecular dynamics simulations, where spatial restraints and a backbone-orientation specified hydrogen-bonding potential are used to guide the MD simulations. For a given target model, FG-MD exploits TM-align[3] to collect the fragments from the PDB library that are structurally similar to the target models. These fragments were used to extract the spatial restraints including Cα distance and contact restraints.

Methods

We submitted five refined models for each target, which were generated using different strategies.

(1) Run the Flexible ModRefiner program for 100 times with different random numbers. Select the model with the lowest ModRefiner energy.

(2) Run the Restrained ModRefiner program once to get the model. Then run FG-MD to further refine the ModRefiner model.

(3) Run the Restrained ModRefiner program for 100 times with different random numbers. Select the model with the lowest energy.

(4) Run the Restrained ModRefiner program for 100 times with different random numbers. Select the model with the lowest MolProbity score.[4]

(5) Run the Flexible ModRefiner program for 100 times with different random numbers. Select the model with the lowest MolProbity score.

When we run the flexible ModRefiner, there is one parameter which controls the weight of the spatial restraints. This parameter is in [0,100]. Since GDT-TS score to the native was given for each CASP10 refinement target, we used the GDT-TS score as the weight in CASP10, which is proportional to the quality of the target reference model.

For the targets where the regions needed to rebuild are informed by the organizers, we kept these regions completely flexible during the ModRefiner energy minimization.

Results

We tested our procedures on the 35 refinement targets from CASP8 and CASP9. It was shown that the five models, generated by the procedures as described above, are generally better than the initial model in most of the features of GDT score and MPscore (see Table 1).

Table 1. Refinement result on 35 CASP8 and CASP9 targets

	GDT-HA	GDT-SC	RMSD	MPscore
Initial	56.330	31.283	3.400 Å	2.627
Model 1	58.096	**33.675**	3.411 Å	2.434
Model 2	**58.195**	31.475	**3.369 Å**	2.524
Model 3	57.817	33.354	3.370 Å	2.434
Model 4	57.661	33.454	3.371 Å	**2.257**
Model 5	57.506	33.264	3.454 Å	2.261

Availability

ModRefiner server and the package are available at http://zhanglab.ccmb.med.umich.edu/ModRefiner/. FG-MD server is at http://zhanglab.ccmb.med.umich.edu/FG-MD/.

1. Xu D, Zhang Y. Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. Biophys J 2011;101(10):2525-2534.
2. Zhang J, Liang Y, Zhang Y. Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling. Structure 2011;19(12):1784-1795.
3. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33(7):2302-2309.
4. Chen VB, Arendall WB, 3rd, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr 2010;66(Pt 1):12-21.

SPARKS-X: Improving the single fold-recognition technique by employing statistical error potentials

Yuedong Yang, Huiying Zhao and Yaoqi 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
Yuedong.yang@gmail.com 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, and SPARKS-X[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

1. Structural model: The query sequence was aligned with pre-compiled structural library, and the template with the highest alignment scores is selected for 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[2].

2. Sequence-based contact prediction: We predict contact map by using 78 features including the direct information[3,4], sequence profile by HHblits[5], and predicted SS, main-chain torsion angles, and ASA. The predictor is trained using libsvm on all 129 CASP9 protein targets.

Availability

The SPARKS-X structure prediction server is available on http://sparks.f3322.org/sparks-x, and the contact prediction server is in built.

1. Yang Y, Faraggi E, Zhao H, Zhou Y. Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. Bioinformatics;27(15):2076-2082.
2. Yang Y, Zhou Y. Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions. Protein Sci 2008;17(7):1212-1219.

3. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C. Protein 3D structure computed from evolutionary sequence variation. PLoS One;6(12):e28766.
4. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci U S A;108(49):E1293-1301.
5. Remmert M, Biegert A, Hauser A, Soding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods;9(2):173-175.

CAD-score: a new method for evaluation of protein structural models

K. Olechnovič and Č. Venclovas

Institute of Biotechnology, Vilnius University
kliment@ibt.lt

The ability to evaluate protein models against the experimentally determined reference structure is crucial for the development and benchmarking of protein structure prediction methods. Although a number of evaluation scores have been proposed to date, many aspects of model assessment still lack desired robustness. To remedy the situation we developed CAD-score [1] (contact area difference score), a new evaluation function quantifying differences between physical contacts in a model and the reference structure.

Methods

The new score uses the concept of residue-residue contact area difference (CAD) introduced by Abagyan & Totrov [2]. Contact areas, the underlying basis of the score, are derived using the Voronoi diagram of spheres that correspond to heavy atoms of van der Waals radii. The Voronoi diagram of spheres is constructed by an algorithm that is especially suited for processing macromolecular structures. The newly introduced CAD-score is a continuous function, confined within fixed limits, free of any arbitrary thresholds or parameters. The built-in logic for treatment of missing residues allows consistent ranking of models of any degree of completeness.

Results

We tested CAD-score on a large set of diverse models and compared it to GDT-TS, a widely accepted measure of model accuracy. Similarly to GDT-TS, CAD-score showed a robust performance on single-domain proteins, but displayed a stronger preference for physically more realistic models. Unlike GDT-TS, the new score revealed a balanced assessment of domain rearrangement, removing the necessity for different treatment of single-domain, multi-domain and multi-subunit structures. Moreover, CAD-score makes it possible to assess the accuracy of inter-domain or inter-subunit interfaces directly. In addition, the approach offers an alternative to the superposition-based model clustering.

Availability

The CAD-score implementation is available both as a web server and a standalone software package at http://www.ibt.lt/bioinformatics/cad-score/.

1. Olechnovič K, Kulberkytė E, Venclovas Č. (2012) CAD-score: A new contact area difference-based function for evaluation of protein structural models. Proteins, doi: 10.1002/prot.24172.
2. Abagyan RA, Totrov MM. Contact area difference (CAD): a robust measure to evaluate accuracy of protein models. J Mol Biol 1997;268(3):678-685.

A t-distribution-based scoring of sequence profile pair in protein distant homology search

M. Margelevičius and Č. Venclovas
Institute of Biotechnology, Vilnius University, Graičiūno 8, LT-02241 Vilnius, Lithuania
minmar@ibt.lt

The concept of homology is at the heart of most studies dealing with protein sequence, structure and function. In the absence of protein structure, inference of homology usually has to rely exclusively on sequence data. At present, most sensitive sequence-based methods use comparison of multiple sequence alignments represented as sequence profiles. Sensitivity of such methods strongly depends on algorithms of profile construction and comparison.

Methods
We propose scoring and comparison of profiles based on statistical theory: The multivariate t-distribution is used to describe the distribution of target profile probabilities. Relating to this type of distribution, we develop a new expression of log-odds scores to score a pair of profiles. To reveal the utility of the new scoring method, we perform a benchmark test on a set of distantly related proteins and compare the results with the existing profile comparison methods by the ROC analysis.

Results
The proposed paradigm of scoring has several important and useful features. By using either the multivariate or matrix-variate t-distribution, the paradigm can be easily extended to the level of profile contexts. Moreover, it can be readily included in Bayesian non parametric statistics. The latter enables statistical clustering of profile segments, thus making profile-pair scores group-oriented and more sensitive.

CAMEO

Juergen Haas, Tobias Schmidt, Andrew Waterhouse, Marco Biasini, Stefan Bienert, Konstantin Arnold, Tiziano Gallo Cassarino, Valentina Romano, Lorenza Bordoli, Torsten Schwede

SIB & Biozentrum University of Basel

CAMEO (http://www.cameo3d.org) is a service for continuous automated model assessment. The first category within the framework assesses protein structure prediction servers. Protein structure modeling is widely used in the life science community to build models for proteins, where no experimental structures are available. However, depending on the specific target protein and the applied modeling approach, the accuracy of computational models may vary significantly between different modeling servers.

CAMEO uses the amino acid sequences of the weekly PDB releases to continuously assess the accuracy and reliability of protein structure modeling servers. Retrospective evaluation of prediction accuracy allows users of models to select the most suitable tool for a given modeling problem.

CAMEO evaluates prediction accuracy, and hence provides an independent blind benchmark to document the performance of new algorithms. Since the accuracy requirements for different scientific applications vary, CAMEO offers a variety of scores assessing different aspects of a prediction (coverage, local accuracy, completeness, etc.) to reflect these requirements.

A second category for continuous assessment are the Ligand Binding Site Residue Predictions, which just has opened, along with the possibility to annotate ligands within CAMEO and thus aid the method developers, which in turn can produce more refined predictions for the user of these services.

CAMEO has been inspired by EVA[1] and LiveBench[2] among others.

1. Eyrich, V.A., Marti-Renom, M.A., Przybylski, D., Madhusudhan, M.S., Fiser, A., Pazos, F., Valencia, A., Sali, A., Rost, B. Bioinformatics (2001) 17(12): 1242-1243.
2. Bujnicki, J. M., Elofsson, A., Fischer, D. and Rychlewski, L. (2001), Protein Science, 10: 352–361.

CAMEO Ligand Binding - Continuous Automated Evaluation of Ligand Binding Site Predictions

Tobias Schmidt, Andrew Waterhouse, Marco Biasini, Stefan Bienert, Konstantin Arnold, Tiziano Gallo Cassarino, Valentina Romano, Lorenza Bordoli, Juergen Haas, Torsten Schwede

SIB & Biozentrum University of Basel, Switzerland

The task of predicting binding sites from a protein's sequence is of high relevance for life science research, ranging from functional characterization of novel proteins to applications in drug design. Consequently, the development of automated methods for predicting ligand-binding sites has received increasing attention over the past years.

In order to help addressing relevant biological questions, the predictions need to be specific and accurate. Thus, in the CAMEO (http://www.cameo3d.org) ligand binding category we continuously assess ligand binding site predictions to evaluate the current state of the art of prediction methods, identify possible bottlenecks, and further stimulate the development of new methods. On average 44 targets with relevant ligands are assessed per week, which allows to draw statistically significant conclusions.

As not all ligands are biologically relevant, CAMEO now features a Structure Annotation system, where ligands are annotated following a classification ontology to distinguish between irrelevant (e.g. buffer, solvent or covalent post translational modification) and relevant ligands (e.g. non-covalently bound natural or synthetic ligands).

The CAMEO framework has been inspired by EVA[1] and Livebench[2] among others.

1. Koh, I. Y., et al. Nucleic acids research 2003, 31(13): 3311-3315
2. Bujnicki J.M., Elofsson A., Fischer D., Rychlewski L. Protein Sci. 2001, 10(2): 352–361.

CAMEO-QE: An automated platform for continuous assessment of local model quality estimation programs

Alessandro Barbato, Juergen Haas,* and Torsten Schwede

Biozentrum, University of Basel, Basel, Switzerland
SIB Swiss Institute of Bioinformatics, Basel, Switzerland
* Presenting author

Model quality evaluation plays a central role in protein structure prediction, and during the last decade, various approaches for estimating the quality of a model (MQAPs) enabling global and/or local (or per-residue) quality estimation have been developed.

According to the CASP9 assessment, the best scoring global MQAPs are nowadays able to perform a nearly optimal relative distinction between "good" and "bad" models, however, local quality estimation still shows significant room for improvement. Given the need of having reliable MQAPs to evaluate *a priori* the usefulness of a model for the biological problem at hand, the community is currently directing its efforts toward the improvement of approaches to estimate the quality of models at a residue level.

Thus, to help both users and MQAP developers we have envisioned an automated way to continuously assess the accuracy of local quality estimation (QE) tools and have added a new category "CAMEO-QE" to the Continuous Automated Model EvaluatiOn (CAMEO [1]) framework, which is inspired by EVA [2] and Livebench [3]. CAMEO so far performs blind assessments of protein structure prediction and ligand binding site prediction methods on the weekly pre-released sequences of the PDB.

As initial proof of concept, we have benchmarked four widely used tools for local model quality evaluation (QMEAN [4], Prosa2003 [5], Dfire [6] and Verify3D [7]) using CAMEO modeling data collected over 1 year, comprising ~11000 models of diverse accuracy.
Residues in the models were classified as "correct" and "incorrect" applying different thresholds for lDDT scores, S-scores and Cα-distances based on least squares superposition relative to the target structures. Based on this assignment, the performance of the various local MQAPs was evaluated through ROC analysis, as this measure does not dependent on the exact nature of the different MQAP functional forms. Here, we present the outcome of this study, including an analysis of differences in accuracy for functionally interesting regions (interface regions, ligand binding sites, etc.) as well as the overall difficulty of the modelling task (easy / hard TBM targets).

1. http://www.cameo3d.org/
2. Eyrich VA, Martí-Renom MA, Przybylski D, Madhusudhan MS, Fiser A, Pazos F, Valencia A, Sali A and Rost B. "EVA: continuous automatic evaluation of protein structure prediction servers.", Bioinformatics. 2001 Dec;17(12):1242-3.
3. Rychlewski L and Fischer D. "LiveBench-8: the large-scale, continuous assessment of automated protein structure prediction.", Protein Sci. 2005 Jan;14(1):240-5.
4. Benkert P, Biasini M and Schwede T. "Toward the estimation of the absolute quality of individual protein structure models.", Bioinformatics. 2011 Feb 1;27(3):343-50.

5. Wiederstein M and Sippl MJ. “ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins.”, Nucleic Acids Res. 2007 Jul;35(Web Server issue):W407-10.
6. Zhang C, Liu S, Zhou H and Zhou Y. “An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state.”, Protein Sci. 2004 Feb;13(2):400-11.
7. Eisenberg D, Lüthy R and Bowie JU. “VERIFY3D: assessment of protein models with three-dimensional profiles.”, Methods Enzymol. 1997;277:396-404.

DisMeta – a Meta Server for Construct Optimization

Y. J. Huang and G. T. Montelione

Center for Biotechnology and Medicine, Rutgers University
Northeast Structural Genomics Consortium
679 Hoes Lane, Piscataway, NJ 07076

yphuang@cabm.rutgers.edu

Natively disordered or unstructured regions in proteins are both common and biologically important, particularly in modulating intermolecular recognition processes. From a practical point of view, however, such disordered regions often can pose significant challenges for crystallization. Disordered regions are also detrimental to NMR spectral quality, complicating the analysis of resonance assignments and three-dimensional protein structures by NMR methods. Identification of such disordered regions, by either experimental or computational methods, is a fundamental step in the NESG (Northeastern Structural Genomics Consortium) structure production pipeline, allowing the rational design of protein constructs that have improved expression, better solubility, improved crystallization, and which provide better quality NMR spectra. The DisMeta Server has been developed by NESG as a construct design and optimization tool.

Methods

The DisMeta Server runs several different disorder prediction software, including DISEMBL(1), DISOPRED2 (2), DISPro (3), FoldIndex (4), GlobPlot2 (5), IUPred (6), RONN (7), VL2 (8). The DisMeta Server also provides sequence-based structural prediction results from other bioinformatics software, including PROF (9), PSIPred (10), SignalP (11), TMHMM (12), Coils (13), SEG (14) and ANCHOR (15). In CASP10, the disorder predictions are calculated based on the Disorder Consensus and SEG results.

Results

The DisMeta results were compared with experimental NMR and HDX-MS data and had a very good agreement in general. We are also using this round of CASP as a performance evaluator.

Availability

The Dismeta server is available at this site: www-nmr.cabm.rutgers.edu/bioinformatics/disorder.

1. Linding, R., Jensen, L. J., Diella, F., Bork, P., Gibson, T. J., and Russell, R. B. (2003) *Structure* **11**, 1453-1459
2. Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F., and Jones, D. T. (2004) *J Mol Biol* **337**, 635-645
3. Cheng J, S. M., Baldi P. . (2005) *Data Mining and Knowledge Discovery* **11**, 213-222
4. Prilusky, J., Felder, C. E., Zeev-Ben-Mordehai, T., Rydberg, E. H., Man, O., Beckmann, J. S., Silman, I., and Sussman, J. L. (2005) *Bioinformatics* **21**, 3435-3438
5. Linding, R., Russell, R. B., Neduva, V., and Gibson, T. J. (2003) *Nucleic Acids Res* **31**,

3701-3708
6. Dosztanyi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005) *J Mol Biol* **347**, 827-839
7. Yang, Z. R., Thomson, R., McNeil, P., and Esnouf, R. M. (2005) *Bioinformatics* **21**, 3369-3376
8. Vucetic, S., Brown, C. J., Dunker, A. K., and Obradovic, Z. (2003) *Proteins* **52**, 573-584
9. Rost, B., Yachdav, G., and Liu, J. (2004) *Nucleic Acids Res* **32**, W321-326
10. Jones, D. T. (1999) *J Mol Biol* **292**, 195-202
11. Emanuelsson, O., Brunak, S., von Heijne, G., and Nielsen, H. (2007) *Nat Protoc* **2**, 953-971
12. Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. (2001) *J Mol Biol* **305**, 567-580
13. Lupas, A., Van Dyke, M., and Stock, J. (1991) *Science* **252**, 1162-1164
14. Wootton, J. C., and Federhen, S. (1996) *Methods Enzymol* **266**, 554-571
15. Meszaros, B., Simon, I., and Dosztanyi, Z. (2009) *PLoS computational biology* **5**, e1000376

EVfold: de novo protein 3D structure from sequence variation

Debora Marks and Chris Sander

Harvard Medical School and Memorial Sloan-Kettering Cancer Center
ecreview@hms.harvard.edu

We use a new method for contact prediction from multiple sequence alignments to compute 3D structure models. The method is based on the well known idea that correlated mutations may be indicative of residue contacts. We address the confounding effect of transitive correlations in chains of pairs, higher order correlations and statistical noise by a maximum entropy approach. The resulting global probability model aims to capture true evolutionary couplings in residue pairs, useful for the discovery of functionally and structurally important interactions.

As a test, we report the accuracy of the method as evaluated for cases of known 3D structures in blinded fashion, both for globular proteins and for alpha-helical trans-membrane proteins. The improvement in prediction accuracy for contacts and for 3D structures compared to earlier methods provides encouragement to apply the method to completely unknown structures. As genuine predictions in unknown territory, we report a set of predicted 3D structures of medically interesting trans-membrane proteins, such as the adiponectin receptor. The current requirement of at least hundreds of sequences in an iso-structural family limits the applicability of the method, e.g., to fewer than 10% of human proteins.

In the future, methods of this type should be increasingly useful, as current sequencing technology will rapidly increase the number of sequences in protein families. The method will be applied in rolling CASP and will be made available on a web server.

1. *First EVfold 3D structures:* DS Marks, LJ Colwell, R Sheridan, TA Hopf, A Pagnani, R Zecchina, C Sander. Protein 3D structure computed from evolutionary sequence variation. *PLoS One* 6, e28766, doi:10.1371/journal.pone.0028766 (2011).
2. *Maximum entropy method development:* F Morcos, A Pagnani, B Lunt, A Bertolino, DS Marks, C Sander, R Zecchina, JN Onuchic, T Hwa, M Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci U S A* 108, E1293-1301, doi:10.1073/pnas.1111471108 (2011).
3. *EVfold_membrane 3D structures*: TA Hopf, LJ Colwell, R Sheridan, B Rost, C Sander, DS Marks. Three-Dimensional Structures of Membrane Proteins from Genomic Sequencing. *Cell*, doi:10.1016/j.cell.2012.04.012 (2012).
4. *Review of co-variation methods*: DS Marks, TA Hopf, C Sander. Predicting Protein Structure from Sequence Variation. *Nature Biotechnology*, 15 Nov 2012

Local Distance Different Test (lDDT): A novel superposition-free similarity measure for protein structures

Valerio Mariani*, Marco Biasini and Torsten Schwede

Biozentrum, University of Basel, Basel, Switzerland
SIB Swiss Institute of Bioinformatics, Basel, Switzerland

Measuring the structural similarity of protein models is central to the CASP experiment. Traditionally, such comparisons are carried out using superposition-based similarity scores (Global Distance Test; GDT) on Cα atoms. GDT quantifies the number of corresponding atoms in two structures that can be superposed within a set of predefined tolerance thresholds. However, being based on rigid-body superposition, they cannot account for changes of relative domain orientation in multi-domain proteins, requiring each domain to be compared separately. The time-consuming domain splitting procedure needs to be carried out manually by the assessors. Additionally, as a Cα measure GDT does not account for accuracy differences of non-Cα atoms, which constitute almost 90% of all atoms in a protein model.

To overcome these limitations, we introduced the local Distance Difference Test (lDDT) score in our assessment of the CASP9 TBM category, which evaluates how well inter-atomic distances in the target protein structures are reproduced in the predicted models. Being superposition-free, the lDDT score can be used to compare multi-domain structures without any prior processing. Furthermore, due to its focus on the conservation of the chemical environment including all atoms, it naturally lends itself to compare functionally relevant regions of the structure (i.e. binding sites or interaction surfaces) locally.

Here, we introduce the improved version of the DDT score, which was applied in the CASP10 assessment: it includes checks of the stereo-chemical quality of the protein structures being compared, and allows the use of multiple reference structures. We show its low sensitivity to domain movements and demonstrate how local DDT scores can be directly used to highlight problematic regions even in multi-domain protein models. We also discuss the significance attached to absolute lDDT score values, and their dependence on the architecture of the proteins being compared. Finally, we show how the use of multiple references removes the need to choose arbitrarily a single reference structure, e.g. in case of assessing against NMR ensembles.

Availability
www.openstructure.org/lddt

Interactive comparative protein structure modeling using Modorama

Jan Kosinski[1,*], Alessandro Barbato[1,4,5*], Pascal Benkert[4,5],
Torsten Schwede[4,5], Anna Tramontano[1,2,3]

1 Department of Physics, Sapienza University P.le Aldo Moro, 5, 00185 Rome, Italy.
2 Center for Life Nano Science @Sapienza, Istituto Italiano di Tecnologia, Sapienza University, P.le Aldo Moro, 5, 00185 Rome, Italy.
3 Istituto Pasteur, Fondazione Cenci Bolognetti, Sapienza University, P.le Aldo Moro, 5, 00185 Rome, Italy
4 Biozentrum, University of Basel, Basel, Switzerland
5 SIB Swiss Institute of Bioinformatics, Basel, Switzerland

Several groups in CASP have shown that expert modelers can often outperform automatic methods by "manually" selecting templates and refining target-template alignments. Moreover, human intervention into modeling process is often necessary in practical biological applications where, for example, templates in a particular functional state must be selected.

To enable more accurate "manual" modeling with less effort, we developed Modorama - an integrated web platform for interactive protein homology modeling and analyzing protein families in general. Using Modorama, protein modelers can perform template search and selection, refine target-template alignments, and build and evaluate models starting from the sequence, or evaluate and refine existing target-template alignments.

Modorama is composed of two interconnected applications: MODexplorer[1] and MODalign[2]. MODexplorer takes as input the target protein sequence and finds structures that could serve as templates for modeling. The best templates and alignments can be selected based on a wide variety of sequence, structural and functional annotations. These annotations include template structural features, sequence conservation, quality assessment scores of the alignments and resulting models, as well as ligand, DNA, and RNA binding sites. After selecting the templates, a structural model can be constructed and evaluated using QMEAN energy function. Optionally, target-template alignments can be manually refined prior to modeling using an interactive alignment editor - MODalign. During the refinement, changes in alignment quality scores are automatically updated and potential errors are automatically detected and highlighted.

1. Kosinski, J., Barbato A. & Tramontano A. MODexplorer: an integrated tool for exploring protein sequence, structure and function relationships (submitted to Oxford Bioinformatics)
2. Barbato, A., Benkert, P., Schwede, T., Tramontano, A. & Kosinski, J. Improving your target-template alignment with MODalign. Bioinformatics 28, 1038–9 (2012).

QMEANbrane – a potential of mean force for membrane proteins

Gabriel Studer, Marco Biasini* and Torsten Schwede

Biozentrum, University of Basel, Basel, Switzerland
SIB Swiss Institute of Bioinformatics, Basel, Switzerland

Membrane proteins are an important class of biological macromolecules involved in many processes of the living cell. They account for one third of genes in the human genome and more than 50% of todays drug targets. The unique physico-chemical properties of biological membranes favor interactions that are energetically discouraged in soluble proteins and vice versa. However, most scoring functions employing potentials of mean force have been trained on soluble proteins. Thus, they perform poorly when applied to membrane proteins.

We have developed QMEANbrane, a parameterization of QMEAN targeted at the quality evaluation of membrane protein models. We combine potentials of mean force, trained on oligomeric membrane protein structures with a per-residue weighting scheme. We show that reliable local quality estimation of these models is possible. Additionally, we argue that the rapidly increasing number of experimentally available membrane protein structures allows for training of membrane-specific potentials of mean force close to statistical saturation.

Davis-QAconsensus / Davis-QAconsensusALL – the baseline quality assessment methods

A.Kryshtafovych, B.Monastyrskyy, K.Fidelis

Protein Structure Prediction Center, Genome Center, University of California, Davis, CA, USA

In CASP9, one of the approaches to assess the effectiveness of QA methods was to compare their performance to that of a naïve predictor, which used a simple clustering technique to calculate quality scores[1]. The assessment showed that the best MQA methods could not outperform the naïve predictor - a rather disappointing result. To estimate performance of quality assessment methods in CASP10, we ran a similar experiment and compared CASP10 MQA results to those of two naïve consensus methods that use identical methodology but different clustering datasets.

Methods

The Davis-QAconsensusALL and Davis-QAconsensus naïve predictors are clustering methods that assign quality score to a model based on the average pair-wise similarity of the model to other models submitted on that target. For ranking, the Davis-QAconsensusALL predictor uses all server models submitted on a target (ca. 300 models per target), while the DavisQA-consensus predictor uses best 150 models according to the Davis-QAconsensusALL estimate (or, alternatively, just 20 models selected by the Prediction Center for the stage1 QA experiment - see predictioncenter.org/casp10/#predictions). Both methods superimpose all models in the input set with each other using the LGA algorithm in the sequence dependent mode (with default parameters). Next, for each model the quality score is calculated by averaging the GDT_TS scores from all pair-wise comparisons in the set, followed by model completeness scaling.

Note that the Davis-QAconsensus method had access to the same information as all registered QA predictors, while Davis-QAconsensusALL method used extra models (usually of poorer quality) for generating its rankings. Even though this may seem as a substantial advantage, the results of the two methods differ only marginally as the final correlation coefficients are calculated on the identical subsets of 150 (or 20) models.

Results

As in CASP9, simple clustering methods proved to be on par with leading quality assessment techniques. In the per-target assessment mode (QA1.1, see [1]), the Davis-QAconsensusALL method reached the weighted mean Pearson's correlation coefficient (wmPMCC) of 0.62 (0.59 for Davis-QAconsensus) in ranking the suggested 150 server models, which is equal to the highest wmPMCC for the participating CASP10 QA groups. Results of both naïve methods appeared to be statistically indistinguishable from those of the twelve top-performing groups. In the QA1.2 mode (models for all targets pooled together), both naïve methods attain PMCC of 0.93 (again, the highest value) and are statistically indistinguishable from the five top-performing groups both according to the correlation-based and ROC-based analyses.

1. Kryshtafovych A, Fidelis K, Tramontano A. (2011) Evaluation of model quality predictions in CASP9. Proteins 79 (S10), 91-106.
2. Zemla A. (2003) LGA: A method for finding 3D similarities in protein structures. Nucleic Acids Res 31 (13), 3370-3374.

Sphere Grinder – estimating similarity of structures on a local scale

P. Lukasiak[1,2], M.Wojciechowski[1,3], T. Ratajczak[1,3], K. Hasinski[1,3], B. Monastyrskyy[3], A. Kryshtafovych[3] and K. Fidelis[3]

1 Institute of Computing Science, Poznan University of Technology,Poznan, Poland
2 Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland
3- Protein Structure Prediction Center, Genome Center, University of California, Davis, CA, USA

The vast majority of software tools to measure similarity between protein models and experimental structures used in CASP and elsewhere are based on a single global superposition or a series of global superpositions (as in the basic CASP score, the GDT_TS[1]). In CASP6, a non-rigid body structure comparison tool and measure – the Descriptor ALignment (DAL)[2,3], which focused on finding correct structural alignments in the sense of Local Descriptors of Protein Structure[4], was introduced. This measure was used for evaluation of model-target fitness in CASP6-8. In CASP9 we suggested a conceptually simpler tool for comparing structures using local superpositions - the Sphere Grinder. This tool was tested by the CASP9 refinement and free modeling assessors, and seemed effective in cases of weak model-target similarity, significant structural shifts, or predictions on multi-domain targets, where it oftentimes helped identify models with better structural characteristics.

Methods

The main idea behind the SphereGrinder approach is to use local structure superpositions calculated within spheres centered on all the CAs to compare model and target structures. RMSDs are then calculated between all atoms falling within each sphere in the target and the corresponding atoms in the model. Sphere radii may be selected by the user. Different models can then be ranked by the percentage of residues for which the local structure (i.e. the structure within a sphere centered on that residue) does not deviate from target by more than a pre-defined RMSD cutoff. Two scores are implemented - an all-atom score showing fitness within local spheres normalized by the percentage of atoms predicted by the model and a raw score ignoring atoms missing in the prediction. Alternatively, to provide a fuller picture, a range of radii may be used to calculate a map of model quality.

Results

In CASP10 we report scores calculated for spheres of a 6A radius and an RMSD cutoff of 2A. These scores were used by the TBM and refinement assessors as integral part of their ranking of models. The scores are reported for each target and each model submitted. While SphereGrinder may be used to report a scalar measure of model quality calculated for a single sphere radius and a single RMSD cutoff, it is probably more effective in visualizing results for a range of sphere sizes and RMSD cutoffs all at once, providing a single glance picture of model quality. These results may be visualized using an interactive Sphere Grinder Viewer, via the CASP10 results page.

1. Zemla A, Venclovas C, Moult J, and Fidelis K. (1999) Processing and analysis of CASP3 protein structure predictions. Proteins Suppl. 3, 22-29.
2. Kryshtafovych A, Milostan M, Szajkowski L, Daniluk P, Fidelis K. (2005) CASP6 data processing and automatic evaluation at the protein structure prediction center. Proteins 61 (S7), 19-23.
3. Kryshtafovych A, Prlic A, Dmytriv Z, Daniluk P, Milostan M, Eyrich V, Hubbard T, Fidelis K. (2007) New tools and expanded data analysis capabilities at the Protein Structure Prediction Center. Proteins 69 (S8), 19-26.
4. Hvidsten TR, Kryshtafovych A, Fidelis K. (2009) Local descriptors of protein structure: A systematic analysis of the sequence-structure relationship in proteins using short- and long-range interactions. Proteins 75(4), 870-884.