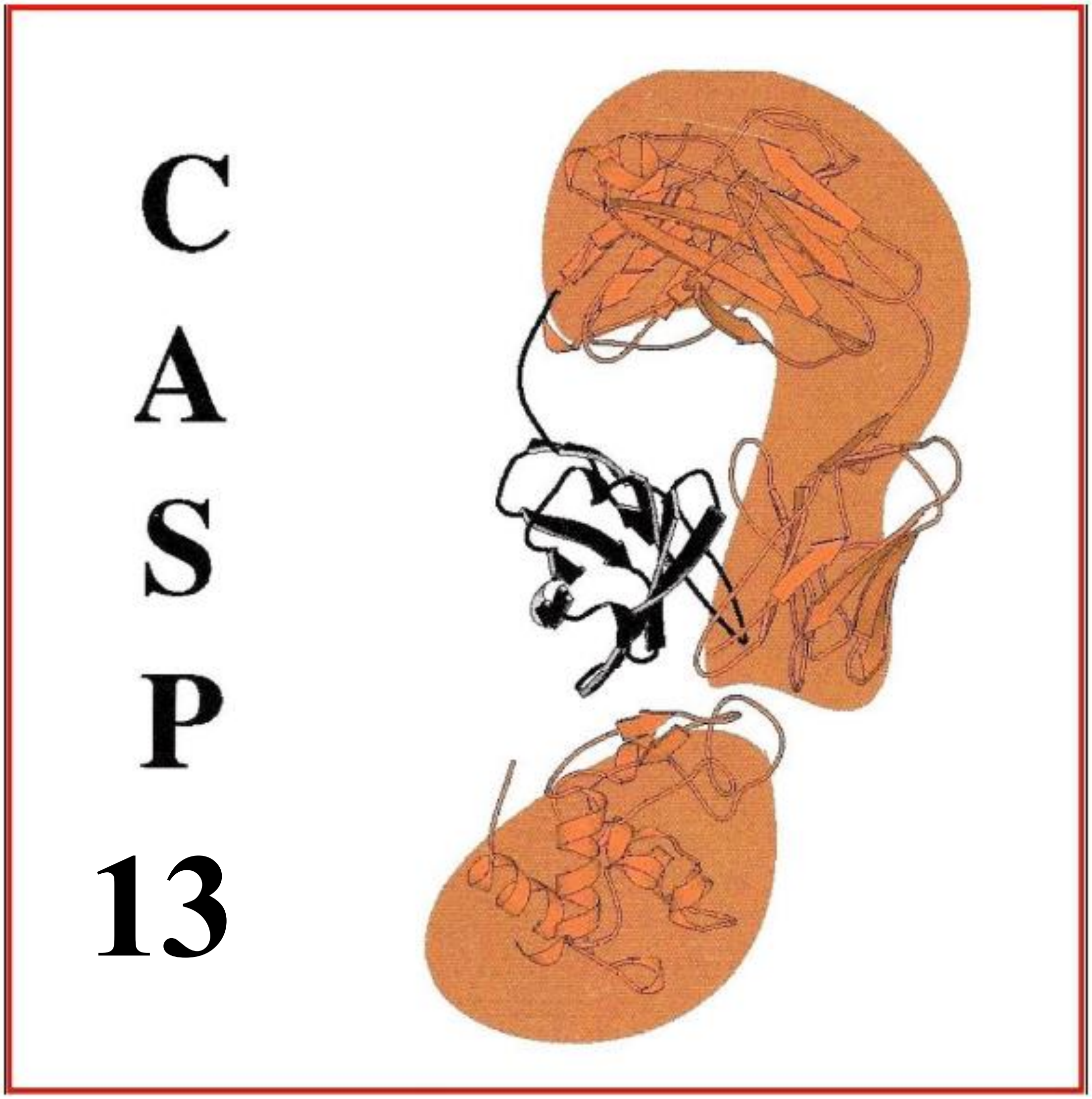
CRITICAL ASSESSMENT OF TECHNIQUES FOR PROTEIN STRUCTURE PREDICTION



Thirteenth meeting
Riviera Maya, Mexico
DECEMBER 1-4, 2018

TABLE OF CONTENTS

Protein model quality assessment using 3D oriented convolutional neural network

G. Pagès[1], B. Charmettant[1] and S. Grudinin[1]

[1] - Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France

Sergei.Grudinin@inria.fr

Protein model quality assessment (QA) is a crucial and yet open problem in structural bioinformatics. The current best methods for single-model QA typically combine results from different approaches, each based on different input features (both structure-based and sequence-based) constructed by experts in the field. Then, the prediction model is trained using a machine-learning algorithm. Recently, with the development of convolutional neural networks (CNN), the training paradigm has been changed. In computer vision, the expert-developed features have been significantly overpassed by automatically trained convolutional filters. This motivated us to apply a three-dimensional (3D) CNN to the problem of protein model QA.

Methods

We present Ornate (Oriented Routed Neural network with Automatic Typing), a novel method for single-model QA. Ornate is a residue-wise scoring function that takes as input 3D density maps. It predicts the local (residue-wise) and the global model quality through a deep 3D CNN. Specifically, the Ornate method aligns the input density maps, constructed from each residue and its neighbourhood, with the backbone topology of the corresponding residue. This circumvents the problem of ambiguous orientations of the initial models[1]. Also, Ornate includes automatic identification of protein atom types.

The input of the network is constituted of 167 density maps, each consisting of 24×24×24 voxels with a 0.8 Å side. Each map represents the density of one type of atoms among the 167 that can be found in proteins. However, such a representation is very sparse. To make the representation dense and reduce the number of network variables, we linearly projected the 167 types into a 15-dimensional space. We wanted to be as rigorous as possible on making assumptions about classifying the atoms. Therefore, we let the network to learn the projection automatically upon training by designing “retyper” projection layer. This followed by three 3D convolutional layers that learn structural features on different scales. Then, two last fully connected layers process the features from the previous layers and output a scalar. We trained the method on structures from the previous CASP experiments using the CAD-score[2] of each residue as the ground truth.

Results

We applied the Ornate method to the QA category of CASP13 as a server. We also tested the performance of this method on CASP 11 and 12 test cases. There, Ornate achieves the state-of-the-art performance for single-model quality assessment when compared to CAD-score as the ground truth. More specifically, we achieved a Pearson correlation of 0.72 and 0.78 on CASP 11 stage 2 and CASP 12 stage 2 datasets, respectively, while Proq3D[3], the best method that we tested, achieved corresponding correlations of 0.72 and 0.80.

Availability

Ornate will be made available on our website at https://team.inria.fr/nano-d/software/Ornate/.

1. Derevyanko,G., Grudinin,S., Bengio,Y., & Lamoureux,G. (2018). Deep convolutional networks for quality assessment of protein folds. *Bioinformatics*, bty494.
2. Olechnovič,K., Kulberkytė,E., & Venclovas,Č. (2013). CAD-score: a new contact area difference-based function for evaluation of protein structural models. *Proteins: Struct., Funct., Bioinf.*, **81**, 149-162.
3. Uziela,K., Menéndez Hurtado,D., Shu,N., Wallner,B., & Elofsson,A. (2017). ProQ3D: improved model quality assessments using deep learning. *Bioinformatics*, **33**, 1578-1580.

Refinement of protein models with additional cross-linking information using the Gaussian network and gradient descent

G. Pagès and S. Grudinin

1 - Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France

Sergei.Grudinin@inria.fr

Cross-linking (XL) experiments provide information on contacts between pairs of residues. As contact prediction methods are becoming more and more reliable, it is interesting to find new ways to integrate contact data into the modeling pipeline of protein structures.

Methods

In order to take advantage of XL data we made a few assumptions. Let us consider two residues (represented by the corresponding alpha carbons) for which the XL experiment has detected a contact. First, we estimated the probability of presence for one alpha carbon with respect to the distance to the second atom. It appeared that we could roughly approximate this probability by a Gaussian distribution centered at zero with the standard deviation specific to each type of XL experiment[1]. We then decided to make a Boltzmann-like hypothesis and considered that there is a pseudo-potential associated to the XL experiment, whose value is given by the logarithm of the presence probability of an alpha carbon. Since we made the hypothesis of a Gaussian distribution of one alpha carbon with respect to the other, this pseudo-potential is a harmonic.

On the next step, we ranked and refined the best models by doing a gradient-based optimization. When moving the model atoms along the row gradient of the XL pseudo-potential, bonds may break, unrealistic local topology may occur, and as a result, the initial secondary structure can also get broken. To preserve the local model topology, the gradient descent step is performed by minimizing the energy of a Gaussian network model, whose equilibrium is the current state, with an additional term from the gradient of the XL pseudo-potential. This is equivalent to solving a linear system of equations. The Gaussian network model, computed by the NOLB library[2], allows large-amplitude realistic motions, with marginal modification of the local topology. However, the accumulation of small perturbations of the local topology during the different steps still may produce unrealistic structures. To tackle this problem, we added to our iterative process an additional minimization of a simple force field with energy terms containing bond length, bond angle, and van der Walls interactions. We continued the refinement until the convergence of the total energy.

Results

This method was applied to all monomeric XL-assisted targets from the CASP13 experiment. We only made a visual inspection of the models during the refinement. We actually expect CASP 13 experiment to provide us valuable assessment on the capability of our method.

Availability

This method will be made available on our website at https://team.inria.fr/nano-d/software/ .

1. Leitner, A., Joachimiak, L. A., Unverdorben, P., Walzthoeni, T., Frydman, J., Förster, F., & Aebersold, R. (2014). Chemical cross-linking/mass spectrometry targeting acidic residues in proteins and protein complexes. Proceedings of the National Academy of Sciences, 111(26), 9455-9460.
2. Hoffmann,A., & Grudinin,S. (2017). NOLB: Nonlinear rigid block normal-mode analysis method. *J. Chem. Theory Comput.*, **13**, 2123-2134.

De novo structure prediction with deep-learning based scoring

R.Evans[*,1], J.Jumper[*,1], J.Kirkpatrick[*,1], L.Sifre[*,1], T.F.G.Green[1], C.Qin[,1], A.Zidek[1], A.Nelson[1], A.Bridgland[1], H.Penedones[1], S.Petersen[1], K.Simonyan[1], D.T.Jones[2] , K.Kavukcuoglu[1], D.Hassabis[1], A.W.Senior[*,1]

** - Equal contribution, 1- DeepMind, London, UK; 2 UCL, London, UK.*

andrewsenior@google.com

A7D CASP13 submissions were produced by three variants of an automatic free-modelling structure prediction system relying on scores computed with deep neural networks. Scoring relied on one of two neural networks: a predictor of inter-residue distances and a direct-scoring network. The basic method used a generative neural network for fragment generation for fragment assembly in memory-augmented simulated annealing. Successive rounds of simulated annealing used fragments from the memory. The third method used full-chain score minimization with gradient descent.

Methods

The systems tested all use multiple sequence alignments (MSA) and profiles generated from HHBlits [2] and PSI-BLAST [3]. No templates were used, nor were server predictions. No manual intervention was made except for domain segmentation of T0999 and final decoy ranking in a handful of cases. In protein complexes, each chain was processed independently.

Scoring

Two neural networks were used for scoring. For the first, a very deep residual convolutional neural network was trained on a non-redundant database of proteins selected from the Protein Data Bank (PDB) to predict the distances between C-beta atoms of different residues, using MSA-based features. With these predictions and a reference distribution, a likelihood score was computed for candidate structures according to the realised distances.

A second deep residual convolutional neural network was trained to directly output a score as a function of structure geometry, MSA-based features and the contact predictions from the first network.

Domain segmentation

Domain segmentation hypotheses for two or three domains were generated by automatic analysis of the full-chain contact matrix prediction derived from the inter-residue distance prediction. Each domain segmentation hypothesis (as well as full chain without segmentation) was folded independently up to eight times with the domains in each hypothesis being folded independently.

Fragment assembly

Two approaches were used for structure modelling. The first was based on fragment assembly. For each domain, a DRAW [4] model of backbone torsion angles, trained on the same PDB subset was sampled to generate a set of overlapping 9-residue fragments. Fragments were inserted with simulated annealing using a score based on our distance predictions for the domain hypothesis plus Rosetta's [1] score2 (Variant 1) or the direct structure scoring without Rosetta (Variant 2).
Repeated rounds of simulated annealing were run, using evolutionary hyper-parameter optimization to tune run-length and start temperature, with successive rounds using fragments from the structures generated in previous rounds.

The best-scoring structures from simulated annealing were relaxed using Rosetta fast relax with our inter-residue distance prediction score and Rosetta's full-atom score.

Domain assembly

After domain-level relaxation, for each domain segmentation, full-chain structures were assembled from domain structures with simulated annealing and further relaxed. The best-scoring full-chain structure for each run of each domain segmentation hypothesis for each method was chosen.

Direct structure optimization

An alternative structure modelling approach was used for Variant 3 without any domain segmentation. Here we used gradient descent of a combination score (inter-residue distance prediction score + neural-network-based torsion angle prediction likelihood + score2) to optimize *full chain* structures, parameterised with torsion angles.

Decoy selection

Five ranked structure predictions were submitted for each "all groups" target. Initial submissions used variants 1 & 2 in parallel, but submissions from T0975 on used variants 1 & 3 in parallel. The 5 candidate submissions were the best scoring from among the independent runs of the two different methods, with a bias towards selecting from variants 2 or 3, and manual ranking in a handful of cases.

Acknowledgements

We are grateful for fruitful discussions with David Silver and Oriol Vinyals and contributions from Steve Crossan, Marek Barwinski, Ruoxi Sun, Carl Elkin, Peter Dolan, Matthew Lai and Yujia Li as well as many others at DeepMind.

We also gratefully acknowledge the use of several tools and datasets (Rosetta, HHBlits, HHPred, PSI-BLAST, PDB, CATH).

1. Das,R., Baker,D. (2008) Macromolecular modeling with rosetta. *Annu Rev Biochem.* 77:363–382.
2. Remmert,M., Biegert,A., Hauser,A. (2012) HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment *Nature Methods*, 9(2), 173-175.
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. Gregor,K., Danihelka,I., Graves,A., Rezende,D.J. (2015) DRAW: A recurrent neural network for image generation *arXiv*:1502.04623

AIR: An artificial intelligence-based protocol for protein structure refinement using multi-objective particle swarm optimization

Di Wang[1,2], Ling Geng[1,2], Yu-Jun Zhao[1,2], Yang Yang[3], and Hong-Bin Shen[1,2]

1 - Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, 2 -Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240, China, 3 Department of Computer Science and Engineering, Shanghai Jiao Tong University

hbshen@sjtu.edu.cn

The AIR is an Artificial Intelligence-based protein Refinement method, which is constructed using a multi-objective particle swarm optimization (PSO) protocol. The basic motivation of AIR is trying to solve the bias problem caused by minimizing only a single energy function due to the significant diversities of different protein structures. Thus, the fundamental idea of our method is to use multiple energy functions as multi-objectives so as to correct the potential inaccuracy from a single function. We designed a multi-objective PSO algorithm-based structure refinement, where in the protocol each protein structure is taken as the particle. The particles will move to better positions during the process of structure refinement. Whether current positions are good, i.e. the quality of current particles (structures), will be evaluated by three energy functions. Then, we will decide which particles are non-dominated particles, which means the value of at least one objective energy for those particles are less than all other particles. These non-dominated particles will be put into a set called Pareto set, which is the collection of global best and local best particles in our refinement iterations. After enough iteration times, the particles from the Pareto set will be ranked and top 5 of them will be outputted, which are the final refined structures.

Methods

Step 1: Initial Particle swarm construction (Structure templates):
In addition to using the model given by the CASP website, we also selected two models from the submitted models predicted by other servers as the initial templates. From these three initial templates, we give each of them some random perturbation, resulting in a total of 50 different particles, each template producing roughly the same number of particles.

Step 2: Particle movement (Structure refinement) in a multi-objective way:
In the second step, we apply PSO[1]algorithm in the evolution. In each iteration, the global best particle and local best particle are selected to guide current particles' direction, with some random disturbance involved. What's more, unlike other algorithms that use a single energy function to evaluate the model, we use three different evaluation methods (Charmm[2], Rosetta[3] and Rwplus[4] scoring function) as the fitness function. By comparing the three scores' dimensions of different particles, the non-dominated relationship builds the Pareto sets[5]. Compared with the single energy function algorithm, this multi-objective optimization algorithm can have advantages of three different energy functions and reduce the risk of inaccuracy from using only a single energy function because there is no energy function that can be suitable to all protein structures' evaluation.

Step 3: Solution ranking:
After enough iteration times, the final Pareto set was considered as the candidate solution sets. The three-dimension structure of the proteins in the final candidate solution set are clustered and the protein structure sets of each class are sorted by the knee algorithm[6]. The top ranked structures are selected.

1. Kennedy J. (2011) Particle swarm optimization. Encyclopedia of machine learning. Springer, Boston, MA.760-766.
2. Brooks,B.R. et al. (1983) CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. Journal of computational chemistry. 4(2): 187-217.
3. Rohl,C.A. et al. (2004) Protein Structure Prediction Using Rosetta. Methods in Enzymology. 383:66-93.
4. Zhang J, Zhang Y . (2010) A Novel Side-Chain Orientation Dependent Potential Derived from Random-Walk Reference State for Protein Fold Selection and Structure Prediction. PLoS ONE 5(10): e15386.
5. Coello,C.A. et al. (2004) Handling multiple objectives with particle swarm optimization. IEEE Transactions on Evolutionary Computation. 8:256-279.
6. Branke,J. et al. (2004) Finding knees in multi-objective optimization. International conference on parallel problem solving from nature. Springer, Berlin, Heidelberg. 722-731.

Protein Contact Prediction with Multi-scale Residual Convolutional Neural Network

Yang Liu, Wei Qian, Yunan Luo, Qing Ye, Jian Peng

University of Illinois, Urbana Champaign

jianpeng@illinois.edu

Substantial progress has been made in protein contact map prediction. However, most current methods for contact map prediction[1-6] may not predict reliable contact pairs when the quality of the input multiple sequence alignment (MSA) is poor. Here we present a new deep convolutional neural network-based contact prediction algorithm, aiming to capture multi-scale patterns in long-range evolutionary couplings and improve the predictive performance.

Methods

Given a protein sequence, our method first used HHblits (with an E-value threshold of 1e-3) to produce a multiple sequence alignment (MSA). If the MSA has fewer than 1,000 hits, we instead utilized JackHMMER (with an E-value threshold of 10) to gather more sequences in the MSA. We then visualized the gap patterns in the MSA and segmented the original sequence into multiple domains if needed. If the sequence is split into multiple domains, we re-ran the previous process for each domain and obtained one MSA for each domain. After generating MSA, we obtained 1D and 2D features for each domain. The 1D features include column-wise amino acid composition from MSA, the secondary structure predicted by PSIPRED[7] and solvent accessibility predicted by SOLVPRED[5]. The 2D features include co-evolutionary patterns, mutual information (MI), normalized MI and the mean contact potential.

Features of domains were concatenated together and used as input to a convolutional neural network.[9] The network consists of multiple residual convolutional blocks. Each residual block contains two Conv-BatchNorm-ReLU layers. Given the input, the network firstly applied ten residual blocks to obtain high-resolution features. Then it applied a max-pooling layer with five stacked residual convolutional blocks to extract higher-level, lower-resolution features. After that, another max-pooling layer and 5 residual blocks are stacked on the top. Finally, all three-resolution features were up-sampling and adding together as the final feature. Another two convolutional layers with kernel size 1 and filter number 32 were applied to generate the final contact prediction.

In our experiments, we tried four different strategies: *BetaContact* utilized HHblits as MSA, *GammaContact* utilized JackHMMER as MSA, *DeltaContact* utilized the combination of HHblits and JackHMMER as MSA. Finally, *AlphaContact* utilized the method we described as the MSA option (If the number of hits in HHblits MSA is smaller than 1,000, we used JackHMMER instead.)

Results

In training, we used a filtered dataset based on the ASTRAL-2.06 database[8]. Two residues are considered as a contact pair if their Cβ-Cβ distance is smaller than a predetermined threshold. We ensemble six models as final predictor, four models are trained with a cutoff 8.0 Å, one with a cutoff 7.5 Å and one with 8.5 Å.

The accuracy of our server is evaluated locally on three datasets including CAMEO, CASP10/11 and CASP12 dataset. To evaluate our method, only pairs of residues with distance smaller than 8.0Å were considered as contact pairs. We compared the mean precision of the top L, L/2, L/5 and L/10 predictions on both medium-range (11<distance<24) and long-range (distance>23) contact pairs. Our deep residual fully convolutional network significantly outperforms CCMPred.

			Medium Range: 11<i - j| <24				Long Range: |i - j|>23			
Dataset	#protein	Method	L	L/2	L/5	L/10	L	L/2	L/5	L/10
CASP10 + CASP 11 Dataset	228	CCMPred	20.31	29.57	44.60	52.95	29.81	39.42	50.56	56.51
		Our Method	37.28	55.92	74.96	83.00	54.73	68.43	77.40	80.84
CASP12 Dataset	25	CCMPred	9.23	13.23	20.60	27.08	14.95	19.33	25.22	30.90
		Our Method	23.95	35.33	47.61	56.25	34.80	43.75	50.88	52.13
CAMEO Dataset	219	CCMPred	13.04	19.10	29.34	36.90	21.59	28.94	38.31	44.50
		Our Method	28.51	43.04	60.70	69.13	45.11	57.28	67.00	71.76

1. Morcos,F., Pagnani,A., Lunt,B., Bertolino,A., Marks,D.S., Sander,C., ... & Weigt,M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci.*, **108**, E1293-E1301.
2. Kamisetty,H., Ovchinnikov,S., & Baker,D. (2013). Assessing the utility of coevolution-based residue–residue contact predictions in a sequence-and structure-rich era. *Proc. Natl. Acad. Sci.* ,**110**, 15674-15679.
3. Seemayer,S., Gruber,M., & Söding,J. (2014). CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. *Bioinformatics*, **30**, 3128-3130.
4. Wang,S., Li,W., Zhang,R., Liu,S., & Xu,J. (2016). CoinFold: a web server for protein contact prediction and contact-assisted protein folding. *Nuc. Acids Res.*, gkw307.
5. Jones,D.T., Singh,T., Kosciolek,T., & Tetchner,S. (2015). MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics*, **31**, 999-1006.
6. 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.
7. McGuffin,L.J., Bryson,K., & Jones,D.T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics*, **16**, 404-405.
8. Fox,N.K., Brenner,S.E., & Chandonia, J. M. (2014). SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nuc. Acids Res.*, **42**, D304-D309.
9. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

AP_1 structure prediction in CASP13

Hyung-Rae Kim

Department of Electrophysics, Kyonggi University, 154-42 Gwanggyosan-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do, South Korea 16227

hr_kim@kgu.ac.kr

AP in AP_1 stands for "Allied Protocols in protein science program suite", which is being prepared for publication. AP_1 is the first ever developed AP program package. Thus, it was named AP_1. The goal of AP_1 is single chain protein structure scoring. AP_1 inherits several characteristics of PRESCO [1], such as database search and structure retrieval without calculating pair-wise potentials and without building a fixed form potential.

Methods

The goal of AP_1 is to accurately score not only the topology of a protein structure, but also the side-chain positions of the high-accuracy template-based models.

The newly incorporated assumption of AP_1 was that "amino acids might contribute differentially to the formation steps of a protein structure". Thus, each contribution of amino acids in structure scoring was employed differentially and implemented. Moreover, we found and employed three new scoring matrices, which had not been used before.

AP_1 was used on CASP12 for the first time. However, owing to the difficulty in balancing between the high accuracy scoring and topology scoring, AP_1 did not perform well for CASP12. We changed the weighting-scheme in CASP13 and tested our AP_refine protocol, which is currently under development.

Thus, our structure prediction pipeline consists of the following points.

1. Five of the best models were picked using AP_1 from all submitted server models of CASP13.
2. Five of the best models were picked and used as the seed model for our refinement protocol.
3. Subsequently, five generated models were added to the seed models.
4. We applied AP_1 again to the above candidate models and selected the five best models to submit.

In CASP13, we submitted 415 models for 83 TS regular targets.

Availability

AP_1 is being prepared for publication. Once published, its standalone executable version would be accessible as an appended material.

1. Kim, H. Kihara, D. (2014) *Proteins: Structure, Function, and Bioinformatics*, **82**: 3255-3272

Template-Guided & Coevolution-Restrained Protein Structure Prediction Using Optimized Folding Landscape Force Fields

M. Chen[1,2], X. Chen[1,3], S. Jin[1,4], C.A. Beunos[1,4], W. Lu[1,5], N.P. Schafer[1], X. Lin[1,5], J.N. Onuchic[1,3,4,5] and P.G. Wolynes[1,3,4,5*]

1 - Center for Theoretical Biological Physics, Rice University; 2 - Department of Bioengineering, Rice University; 3 - Department of Chemistry, Rice University; 4 - Department of Biosciences, Rice University; 5 - Department of Physics, Rice University

*pwolynes@rice.edu

That protein polymers gold spontaneously to organized structures is a remarkable physical phenomenon. The essential paradoxes of how proteins are able to fold have been successfully resolved by energy landscape theory.[1] The fact that most globular proteins have a definite structure that is kinetically accessible indicates that the energy landscapes of globular proteins are largely funneled towards their native states, a fact that has come to be known as the Principle of Minimal Frustration.[2] Distilled to mathematical form, this principle states that the native interactions are on average stronger than the possible non- native interactions, which leads to a funneled folding landscape for the protein that guides the protein's Brownian motions towards native-like configurations without encountering many metastable traps.[2] The necessary mathematical framework based on an analogy to spin glasses and a machine learning formalism quantifies the concept of minimal frustration and thereby also provide algorithms to learn and to optimize structure prediction force fields[3]. Based on the energy landscape theory of protein folding, we optimized the Associative memory, Water mediated, Structure and Energy Model (AWSEM), whose transferrable potentials have been proven successful in moderate resolution structure prediction[4]. Albeit the success in structure prediction using the physics-based AWSEM forcefield, the most practical method of protein structure prediction today still relies on evolution: the construction of template-based models[5] and extraction of likely contacts between pairs of residues from protein families[6]. In this current study, we combine the template guidance and inferred coevolutionary information with the AWSEM force field, to achieve improved structure prediction. This combined protocol, AWSEM-Suite, yields significant improvement in the quality of protein structure prediction.

Methods

Structure prediction simulations were performed using the Large-Scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) software package. For each protein target, 20 annealing simulations were carried out from 600 K to 200 K in 4 million steps. All the predictions start from a fully extended configuration that was built using PyMol (www.pymol.org). The structures with the lowest AWSEM energy from each of the 20 simulated annealing trajectories were clustered based on their mutual-Q values. High average mutual-Q values within a cluster indicates strong mutual structural similarity. The centroid structure of the cluster with the highest five average mutual-Q was selected as the final candidates for all-atom reconstructions and final submissions.

1: Obtaining tertiary guidance from templates by HHpred.

We used HHpred[7] to find templates for each target using a minimum threshold confidence score of 95. The aligned regions from the templates were selected and renumbered according to their alignment to the target sequence from HHpred. Then, a pairwise distance matrix was created and the entries in the matrix were used to guide folding along the collective variable Q.

2: Contact Restraints Inferred by RaptorX.

The predicted contacts that we use in this study were obtained using the RaptorX-Contact web server[8] with default settings. The threshold confidence score of 0.5 was recognized as true positive contacts in AWSEM-Suite. For some targets, there might be no positive contacts.

3: Rebuilding all-atom models based on coarse-grained predictions using MODELLER.

Since the structures produced by the coarse-grained AWSEM simulation have only backbone and C_β atoms, rebuilding of the side-chains on these coarse-grained structures was performed using MODELLER[9].

Availability

The source code for the AWSEM-Suite forcefield within the LAMMPS suite is available for download on Github (https://github.com/adavtyan/awsemmd). Other documentations and references can be found on this website: http://awsem-md.org.

1. Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. Funnels, pathways, and the energy landscape of protein folding: a synthesis. Proteins 1995, 21, 167–195.
2. Ferreiro, D. U.; Komives, E. A.; Wolynes, P. G. Frustration in biomolecules. Quarterly Reviews of Biophysics 2014, 47, 285–363.
3. Schafer, N. P.; Kim, B. L.; Zheng, W.; Wolynes, P. G. Learning to fold proteins using energy landscape theory. Israel Journal of Chemistry 2014, 54, 1311–1337.
4. Davtyan, A.; Schafer, N. P.; Zheng, W.; Clementi, C.; Wolynes, P. G.; Papoian, G. A. AWSEM-MD: Protein Structure Prediction Using Coarse-Grained Physical Potentials and Bioinformatically Based Local Structure Biasing. The Journal of Physical Chemistry B 2012, 116, 8494–8503.
5. Chen, M., Lin, X., Lu, W., Schafer, N., Onuchic, J. N., & Wolynes, P. G. Template-Guided Structure Prediction and Refinement with an Optimized Folding Landscape Forcefield. Accepted for publication at Journal of Chemical Theory and Computation.
6. Sirovetz, B. J.; Schafer, N. P.; Wolynes, P. G. Protein structure prediction: making AWSEM AWSEM-ER by adding evolutionary restraints. Proteins 2017, 85, 2127– 2142.
7. Soding, J.; Biegert, A.; Lupas, A. N. The HHpred interactive server for protein homol- ogy detection and structure prediction. Nucleic Acids Research 2005, 33, W244–W248.
8. Wang, Sheng, Siqi Sun, Zhen Li, Renyu Zhang, and Jinbo Xu. "Accurate de novo prediction of protein contact map by ultra-deep learning model." PLoS computational biology 13, no. 1 (2017): e1005324.

9: Sali, A.; Blundell, T. L. Comparative Protein Modelling by Satisfaction of Spatial Restraints. Journal of Molecular Biology 1993, 234, 779–815.

Addressing medium-resolution refinement challenges using Rosetta in CASP13

H. Park[1], G.R. Lee[1], and D. Baker[1,2]

1 - Department of Biochemistry and Institute for Protein Design, University of Washington, WA, USA; 2 - Howard Hughes Medical Institute

dabaker@uw.edu

We have shown in our previous study1,2 that low-resolution homology models can be significantly refined, especially when it is small-sized, through iterative applications of model hybridizations3 guided by an advanced implicit solvent energy model in Rosetta4. Nevertheless, we have also identified in CASP12 that this approach could often hurt than refine the starting models when they are of medium accuracy and larger than 120 residues; we hypothesized that the approach was not optimal for such types of problems and can be further improved.

Methods

In CASP13, we attempted to address this issue by adapting the method to be more conservative, by utilizing "annealing" of the restraints from starting model. Here, 10 regular iterations of model hybridizations are followed by 10 annealing iterations at which subset of coordinate restraints derived from superimposed starting model are applied with the weight gradually increasing as iteration goes; the subset of restraints are selected in an ambiguous fashion5,6 to allow local structure reconstructions occur without penalty. The adaptive version was applied to medium-accuracy starting models with GDT-HA from 50 to 70, while the original strategy to the rest having GDT-HA less than 50 (no starting model had GDT-HA over 70 in this CASP). In both strategies, once iterative modeling is done, a representative model is built by averaging the structures within refinement trajectory similar to the lowest energy model, followed by MD simulations-based refinement7. This model was submitted as model1; the rest of the models were selected from the last iteration pool.

Our predictions were submitted by two groups. A group of predictions submitted as "BAKER_AUTOREFINE", were generated by the automated pipeline described above. Of 31 refinement targets, 20 were modeled through adaptive approach, and the rest through low-resolution strategy. For the second group of predictions submitted as "BAKER", human interventions were made to the automatic submissions to further detect regions to reconstruct, to alter refinement strategy, and to utilize co-evolution information if available.

We also explored symmetric refinement of several selected targets. Rosetta symmetry modeling machinery used for homo-oligomer comparative modeling was readily incorporated into refinement pipeline. Symmetric refinement was applied to the targets if heavily intertwined oligomeric structure is observed from reliable template protein structure(s); 3 targets were selected for automatic predictions (R0977-D4, R0979, and R0981-D4), and 2 more for human-guided predictions (R0981-D5 and R0989-D1).

1. Park H., Ovchinnikov S., Kim D.E., DiMaio F. & Baker D. (2018). Protein homology model refinement by large-scale energy optimization. *Proc. Natl. Acad. Sci. U. S. A.* 115 3054–3059.
2. Ovchinnikov S., Park H., Kim D.E., DiMaio F. & Baker D. (2018). Protein structure prediction using Rosetta in CASP12. *Proteins* 86 Suppl 1 113–121.
3. Song Y., DiMaio F., Wang R.Y.-R., Kim D.E., Miles C., Brunette T., Thompson J. & Baker D. (2013). High-resolution comparative modeling with RosettaCM. *Structure* 21 1735–1742.

4. Park H., Bradley P., Greisen P. Jr, Liu Y., Mulligan V.K., Kim D.E., Baker D. & DiMaio F. (2016). Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. *J. Chem. Theory Comput.* 12, 6201–6212.
5. MacCallum J.L., Perez A. & Dill K. (2015). Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference. *Proc. Natl. Acad. Sci. U.S.A.* 112, 6985-6990.
6. Lee G.R., Heo L. & Seok C. (2018). Simultaneous refinement of inaccurate local regions and overall structure in the CASP12 protein model refinement experiment. *Proteins.* 86 (S1), 168-176.
7. Mirjalili V. & Feig M. (2013). Protein Structure Refinement through Structure Selection and Averaging from Molecular Dynamics Ensembles. *J. Chem. Theory Comput.* 9, 1294–1303.

Improving Robetta by a broad usage of sequence data and coevolutionary restraints

I. Anishchenko[1], D.E. Kim[1,2], H. Park[1], Q. Cong[1], G.R. Lee[1] and D. Baker[1,2]

1 - Department of Biochemistry and Institute for Protein Design, University of Washington, WA, USA; 2 - Howard Hughes Medical Institute

dabaker@uw.edu

Robetta[1] (http://robetta.bakerlab.org) is a fully automated structure prediction server which has been continuously ranked as the top performing method in the structure prediction evaluation project, CAMEO (http://www.cameo3d.org), during the past several years. Following our recent success in large-scale structure determination in Rosetta using coevolutionary information[2], we focused on a broader incorporation of GREMLIN-derived constraints[3] into the modeling pipeline, as well as making use of the most recent sequence data to facilitate template detection and extend the template database by new models of Pfam families[4] not represented in the PDB.

Methods

Ultimate database of protein sequences. Several parts of the Robetta modeling pipeline (e.g. templates detection, domain boundaries prediction, coevolutionary constraints generation) are dependent on the availability of homologous sequences to construct multiple sequence alignments for the query protein. Since there is no unified repository for the sequence data, merging various sources into a single database was one of our goals for improving Robetta. We collected sequences from the following resources: (a) UniRef100, (b) NCBI TSA (2616 sets), (c) JGI Metagenomes (7835 sets), and Metatranscriptomes (2623 sets), and Eukaryotes (891 genomes), (d) genomes collected from various genomic center and online depositories (2815 genomes). After merging and removing 100% redundant entries, we ended up with a database of 7B sequences (or 1.5TB in the FASTA format).

Updated database of templates. Using the above sequence database, we enriched HHsearch[5] template profiles that were not diverse (*hhmake* Neff < 7.0) with additional sequences until Neff reached a cut-off value of 11.0. This step was carried out using *hmmsearch* from the HHMER suite[6]. In addition, following the coevolution-based structure determination method developed in Ovchinnikov et al.[2], we built reliable models for an additional ~1,500 families with no known experimental structure. These models were included in the Robetta templates database and were treated in the same way as regular templates derived from the PDB.

Template detection by map_align. The template detection methods used in Robetta (HHSearch[5], Sparks[7], and RaptorX[8]) were supplemented by the *map_align* algorithm which detects partial threads by matching predicted contacts with the contact patterns of known protein structures by an iterative double-dynamic programing approach[2]. *map_align* is only applied to difficult targets where the other three template detection methods do not agree well on the set of partial threads and enough sequence homologs exist to produce reliable GREMLIN contacts ($N_f > 32$).

Higher-order protein features from coevolution data. Rosetta *de novo* structure prediction guided by coevolutionary-derived residue pair constraints proved to be a reliable tool for modeling difficult targets which lack homologs in the PDB, given that diverse enough multiple sequence alignments can be built[2,9]. However, targets with complicated topologies are often hard to build *de novo* even with reliable constraints, so having additional ways of biasing sampling in relevant conformational regions would be desirable. To partially tackle this problem, we incorporated predicted β-strand pairings from *bbcontacts*[10] into the modeling pipeline by using the Rosetta 'jumping' protocol to sample the predicted nonlocal pairings[11].

Structure modeling. The overall Robetta modeling pipeline did not change significantly since previous CASP12[12]. In brief, the initial step is domain boundary prediction, which consists of an iterative search for PDB templates with optimal sequence similarity and structural coverage to the target using the three template detection methods (HHSearch[5], Sparks[7], and RaptorX[8]). For each predicted domain, models are generated using RosettaCM [13]. If enough sequence data exists to accurately predict co-evolving residue-residue pairs, the clusters are re-ranked using this information, and the RosettaCM spatial restraints are supplemented with the predicted contacts. For difficult domains, models are also generated using the Rosetta fragment assembly methodology[14] (RosettaAB), and if GREMLIN contacts are predicted, they are used as restraints for sampling and refinement. Large scale sampling is achieved using the distributed computing project, Rosetta@home (http://boinc.bakerlab.org/). All models are refined using a relax protocol[15] using the latest Rosetta all-atom energy function[16]. For difficult domains less than 150 residues, an iterative hybridization method was used for further refinement which uses RosettaCM and RosettaAB models as input and outputs a single refined model[17]. The RosettaCM and RosettaAB top scoring cluster representatives and the iterative hybridization model are ranked using ProQ2 for the final 5 selected models. Multi-domain targets are assembled into a single model using Rosetta's domain assembly method[18].

Availability

Robetta is available for non-commercial use at http://robetta.bakerlab.org. The Rosetta software suite can be downloaded from http://www.rosettacommons.org. GREMLIN is available for non-commercial use at http://gremlin.bakerlab.org, and map_align can be freely accessed at https://github.com/sokrypton/map_align.

Acknowledgements

We thank Natalia Ivanova, Igor Grigoriev (JGI) and James Thomas (UW) for providing access to the sequence data.

1. Kim, D.E., Chivian, D. & Baker, D. (2004). Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res.* **32**, W526–31
2. Ovchinnikov, S., Park, H., Varghese, N., Huang, P.-S., Pavlopoulos, G.A., Kim, D.E., Kamisetty, H., Kyrpides, N.C. & Baker, D. (2017). Protein structure determination using metagenome sequence data. *Science* **355**, 294–298
3. Kamisetty, H., Ovchinnikov, S. & Baker, D. (2013). Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proceedings of the National Academy of Sciences* **110**, 15674–15679
4. Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G.A., Tate, J. & Bateman, A. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44**, D279–85
5. Söding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**, 951–960
6. Eddy, S.R. (2009). A new generation of homology search tools based on probabilistic inference. *Genome Inform.* **23**, 205–211
7. 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 query and corresponding native properties of templates. *Bioinformatics* **27**, 2076–2082
8. Peng, J. & Xu, J. (2011). RaptorX: exploiting structure information for protein alignment by statistical inference. *Proteins* **79 Suppl 10**, 161–171
9. Ovchinnikov, S., Kinch, L., Park, H., Liao, Y., Pei, J., Kim, D.E., Kamisetty, H., Grishin, N.V. & Baker, D.

(2015). Large-scale determination of previously unsolved protein structures using evolutionary information. *Elife* **4**, e09248

10. Andreani, J. & Söding, J. (2015). bbcontacts: prediction of β-strand pairing from direct coupling patterns. *Bioinformatics* **31**, 1729–1737
11. Bradley, P. & Baker, D. (2006). Improved beta-protein structure prediction by multilevel optimization of nonlocal strand pairings and local backbone conformation. *Proteins* **65**, 922–929
12. Ovchinnikov, S., Park, H., Kim, D.E., DiMaio, F. & Baker, D. (2018). Protein structure prediction using Rosetta in CASP12. *Proteins* **86 Suppl 1**, 113–121
13. Song, Y., DiMaio, F., Wang, R.Y.-R., Kim, D., Miles, C., Brunette, T., Thompson, J. & Baker, D. (2013). High-resolution comparative modeling with RosettaCM. *Structure* **21**, 1735–1742
14. 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., 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 Enzymol.* **487**, 545–574
15. Conway, P., Tyka, M.D., DiMaio, F., Konerding, D.E. & Baker, D. (2014). Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein Sci.* **23**, 47–55
16. Park, H., Bradley, P., Greisen, P., Jr, Liu, Y., Mulligan, V.K., Kim, D.E., Baker, D. & DiMaio, F. (2016). Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. *J. Chem. Theory Comput.* **12**, 6201–6212
17. Park, H., Ovchinnikov, S., Kim, D.E., DiMaio, F. & Baker, D. (2018). Protein homology model refinement by large-scale energy optimization. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 3054–3059
18. Wollacott, A.M., Zanghellini, A., Murphy, P. & Baker, D. (2007). Prediction of structures of multidomain proteins from structures of the individual domains. *Protein Sci.* **16**, 165–175

Protein model construction and docking using particle swarm optimization

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

Biomolecular Modelling Laboratory, The Francis Crick Institute, 1 Midland Road, London NW1 1AT, UK

raphael.chaleil@crick.ac.uk

The construction, optimization and docking of protein models remains challenging. All require extensive sampling of the high dimensional conformational space, which is intractable with methods based on exhaustive enumeration of all possible solutions. In order to address this problem, we have developed a series of heuristic methods based on Particle Swarm Optimization (PSO) to elevate the problem and be able to generate accuracy solutions.

Methods

Our general methodology for fold construction and docking can be described as follows:

i) Fold construction using our automatic server 3D-Jigsaw-SL

The protocol first searches for homologous sequences to the query sequence using HHBlits[1] against a sequence profile database of known structures clustered at 70% sequence identity. A linear *ab initio* polypeptide corresponding to the query sequence is constructed, taking into account the bond lengths, angles and torsion angles accordingly to identified homologous fragments. All the coil regions that are not matched with a structural template are automatically adjusted in torsion angle space. The central core of the algorithm is a constricted PSO[2], which searches for a minimal Dfire[3] statistical pair potential energy. When distance information was available, either from PSICOV[4] or from discontinuous templates, a hookean force was applied as a distance restraint mechanism. Two strategies were applied for folding the structures, the first one adjusts all the torsion angles between all the fragments at once, whereas the second one adjusts the torsion of each linker region (i.e. regions between fragments from templates) one at a time, starting from the N-terminal. The latter technique is computationally more expensive, however, it achieves to generate structures with a smaller radius of gyration (i.e. the structures are more globular). This property allows to generate better, i.e. biophysically sound, models. Finally, the top 10 ranking models from 100 replicates of the algorithm at 10000 iterations (according to Dfire) are then minimized with CHARMM[5] version 22 and the five top structures with best CHARMM energy after minimization are selected for submission.

ii) Docking using SwarmDock

For all predicted homo-oligomeric structures we used a modification to our binary protein-docking algorithm SwarmDock[6]. Our method uses the principles of PSO to search the parameter docking space. The innovation with the new algorithm is to treat each particle within the swarm as an instance of a packed homo-oligomer, constrained by the appropriate symmetry operators. The objective is to optimize the particle space in order to find the most energetically favorable homo-oligomer. Particles move through a multi-parameter space by the optimization of two sets of parameters: orientations and translations of the monomeric units relative to the imposed symmetry and a linear combinations of normal modes that adjust the conformation of each monomer, in the presence of the other monomers, in this simultaneous docking process. For hetero-oligomeric structures we employed our standard SwarmDock protocol[6]. The monomeric models used for docking were selected from the CASP13 server tarballs.

Availability

3D-Jigsaw server (fold construction): https://bmm.crick.ac.uk/~svc-bmm-3djigsaw/SwarmLoop/
SwarmDock server (protein docking): https://bmm.crick.ac.uk/~svc-bmm-swarmdock/

1. Remmert M., Biegert A., Hauser A. & Söding J. (2011). HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment**.** *Nat. Methods*. 9(2),173-5.
2. Eberhart, R. C. & Kennedy, J. (1995). A new optimizer using particle swarm theory. In *Proceedings of the sixth international symposium on micro machine and human science* (pp. 39–43), Nagoya, Japan. Piscataway: IEEE.
3. Y. Yang & Y. Zhou. (2008). Specific interactions for *ab initio* folding of protein terminal regions with secondary structures. *Proteins* 72, 793-803.
4. Jones DT, Buchan DW, Cozzetto D & Pontil M. (2012). PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*. 28(2), 184-90.
5. Brooks BR, Brooks CL 3rd, Mackerell AD Jr, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor RW, Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, WoodcockHL, Wu X, Yang W, York DM & Karplus M. (2009). CHARMM: the biomolecular simulation program. *J. Comput. Chem.* 30(10), 1545-614.
6. Torchala M., Moal I.H., Chaleil R.A.G, Fernandez-Recio, J. & Bates P.A. (2013). SwarmDock: a server for flexible protein-protein docking. *Bioinformatics.* 29(6), 807-9.

BCL::Fold de-novo protein structure prediction through assembly of secondary structure elements followed by molecular dynamics refinement

J.L. Mendenhall, B.P. Brown and J. Meiler

Department of Chemistry, Vanderbilt University

jeffrey.l.mendenhall@vanderbilt.edu

BCL::Fold[1] enumerates and scores possible conformations of a given protein sequence by assembling predicted secondary structure elements (SSEs) in Euclidean space. Sampled conformations are scored using knowledge-based potentials to estimate the free energy difference between sampled conformations. For selected models, loops are added using a cyclic coordinate descent (CCD) algorithm[2]. Afterwards, the selected structures are simulated in the AMBER16 forcefield to optimize packing and pair interactions[3].

Methods

BCL::Fold assembles models from disconnected SSEs. The SSE definitions are input to the folding algorithm from secondary structure prediction algorithms including PSIPRED [4], MASP [5], and OCTOPUS [6]. A Monte Carlo (MC) sampling algorithm is used to create different arrangements of SSEs by sampling SSE moves. Such moves include adding a SSE to the model; translating, rotating, swapping or flipping SSEs of the model; and altering groups of SSEs to create larger domains, such as sheets. The energy function used in the MC algorithm consists of scoring terms including amino acid exposure, contact-order, SSE packing, loop closure and radius of gyration.

We have recently developed an improved amino-acid scoring potential with side-chain orientation term similar to RWPlus[7], but further parameterized to consider the SSE-types of the contacting residues. Further, we have added an improved contact order scoring term that considers the number and type of SSEs between the two SSEs that are in contact.

The BCL was used to fold 20,000 models for each protein target. For small targets (<150 residues), clustering was performed based on RMSD, and the best scoring member from each of the 30 clusters was visually inspected. For larger proteins, the folding simulations produced diverse topologies. Therefore, we visually culled from the top 30 models by score. Between 4-8 models (depending on SSE content) were chosen for loop building using cyclic coordinate descent, followed by molecular dynamics (MD) refinement using Amber16.

MD refinement was conducted using Amber16 with explicit solvation inTIP4P-EW water [8], hydrogen-mass repartitioning [9], and 3 fs timestep. An initial heating phase from 0-300K was conducted with 1 fs timestep. Simulations were run for 100 – 1000 ns depending on target size and the availability of cluster resources. The model with the smallest radius of gyration from the last 30 ns of simulation was taken as the representative of each trajectory. The protein with the lowest average RMSF over the last 30 ns of the simulation was generally taken as the best.

Because the BCL is a de-novo folding algorithm, we did not predict targets with, e.g. > 75% coverage and > 99% confidence for the best template as assessed by the Phyre2 webserver [10], or other indications that the target would benefit greatly from use of templates.

To focus on testing the improvements we have made to the core BCL folding and scoring algorithm, we did not use predicted contacts.

For the SAXS-assisted targets, de-novo models were generated as described above. The top 4 models by BCL::Score that had a BCL::SAXS RMSD[11] in the top 10th percentile across all the decoys were subsequently refined in MD as described above, and the best model by BCL::SAXS-RMSD was selected as the first model for submission. For the homology targets for which SAXS data were provided (S0985 and S0999), the server model with the lowest BCL::SAXS-RMSD was submitted.

Availability

The BCL software suite is available at http://www.meilerlab.org/bclcommons under academic and business site licenses. The BCL source code is published under the BCL license and is available at http://www.meilerlab.org/servers/bcl-academic-license.

1. Karakas, M. et al. BCL::Fold--de novo prediction of complex and large protein topologies by assembly of secondary structure elements. *PLOS ONE* **7**, e49240, doi:10.1371/journal.pone.0049240 (2012).
2. Canutescu, A. A. & Dunbrack, R. L., Jr. Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Sci* **12**, 963-972, doi:10.1110/ps.0242703 (2003).
3. Case, D. A. et al. The Amber biomolecular simulation programs. *J Comput Chem* **26**, 1668-1688, doi:10.1002/jcc.20290 (2005).
4. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* **292**, 195-202 (1999).
5. Mendenhall, J. & Meiler, J. Prediction of Transmembrane Proteins and Regions using Fourier Spectral Analysis and Advancements in Machine Learning., <https://www.doi.org/10.13140/RG.2.1.2545.8724> (2014).
6. Viklund, H. & Elofsson, A. OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. *Bioinformatics* **24**, 1662-1668, doi:10.1093/bioinformatics/btn221 (2008).
7. Zhang, J. & Zhang, Y. A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. *PLoS One* **5**, e15386, doi:10.1371/journal.pone.0015386 (2010).
8. Horn, H. W. et al. Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. *J Chem Phys* **120**, 9665-9678, doi:10.1063/1.1683075 (2004).
9. Hopkins, C. W., Le Grand, S., Walker, R. C. & Roitberg, A. E. Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning. *J Chem Theory Comput* **11**, 1864-1874, doi:10.1021/ct5010406 (2015).
10. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* **10**, 845-858, doi:10.1038/nprot.2015.053 (2015).
11. Putnam, D. K., Weiner, B. E., Woetzel, N., Lowe, E. W., Jr. & Meiler, J. BCL::SAXS: GPU accelerated Debye method for computation of small angle X-ray scattering profiles. *Proteins* **83**, 1500-1512, doi:10.1002/prot.24838 (2015).

Protein structure prediction and refinement by Bhattacharya human group in CASP13

Debswapna Bhattacharya, Rahul Alapati, Md Hossain Shuvo

Department of Computer Science and Software Engineering, Auburn University, Auburn, AL 36849, USA.

bhattacharyad@auburn.edu

We participated in CASP13 tertiary structure prediction and refinement experiments as human group "Bhattacharya", with our newly developed scoreD[1] method for estimating GDT-TS and GDT-HA using deep discriminative binary classifier ensemble, multi-model QA method clustQ[2] based on weighted internal distance comparisons, and novel structure refinement method refineD[3] using machine learning guided restrained relaxation. One purpose of our participation in the human tertiary structure prediction section is to evaluate the integration of single- and multi-model QA methods for model selection and the use of cumulative multi-resolution probabilistic restraints for conservative yet consistent refinement of selected models. We tested an adventurous refinement strategy in the human refinement category by applying multi-resolution probabilistic restraints in a non-cumulative manner.

Methods

We first selected five models from the whole set of structure models generated by the CASP servers using a combination of scoreD and clustQ method, which included the top two models selected by scoreD targeted at modeling GDT-TS (see Bhattacharya-SingQ QA abstract), top two models selected by scoreD targeted at modeling GDT-HA (see Bhattacharya-Server QA abstract), and top one model selected by clustQ (see Bhattacharya-ClustQ QA abstract). In case the estimated score of the top model selected by clustQ is more than 0.5, we ranked it as the top. Otherwise, the top ranked model was the highest scoring model by scoreD targeted at modeling GDT-TS, while the clustQ selection ranked as fifth. For each of the top five models, four sets of multi-resolution restraints (0.5, 1, 2, and 4Å) centered on the C_α atom of each residue were simultaneously applied in a cumulative manner for all residues weighted according to their probabilities as predicted by the trained binary classifier ensemble. We subsequently employed restrained relaxation protocol for 25 iterations using four parallel threads to generate a total of 100 refined models and used probabilistic combination of the binary classifiers to select the highest scoring model to be submitted.

Structure refinement protocol of the Bhattacharya human group in CASP13 is based on refineD pipeline, that is identical to that used in the Bhattacharya-Server group participating in the refinement category (see Bhattacharya-Server TR abstract), except that instead of selecting the top five refined models from amongst twenty refined models, we generated 100 refined models by employing restrained relaxation for 25 iterations using four parallel threads with each parallel thread using a different restraint resolution and the final five submitted refined structures were selected from the pool of 100 refined models via probabilistic combination of the binary classifiers.

Results

In Figure 1, we present the average GDT-TS score of the first submitted model for all human and server predictors participating the CASP13 tertiary structure prediction experiment for 11 "all groups" targets that could be identified in PDB as of writing this abstract (T0953s1, T0953s2, T0954, T0955, T0958, T0960, T0963, T0965, T0966, T1009, T1016). It shows that Bhattacharya human group (390) ranks at the 11th position with an average GDT-TS score of 52.94, while the highest average GDT-TS score is achieved by predictor 089 with an average GDT-TS score of 55.13.

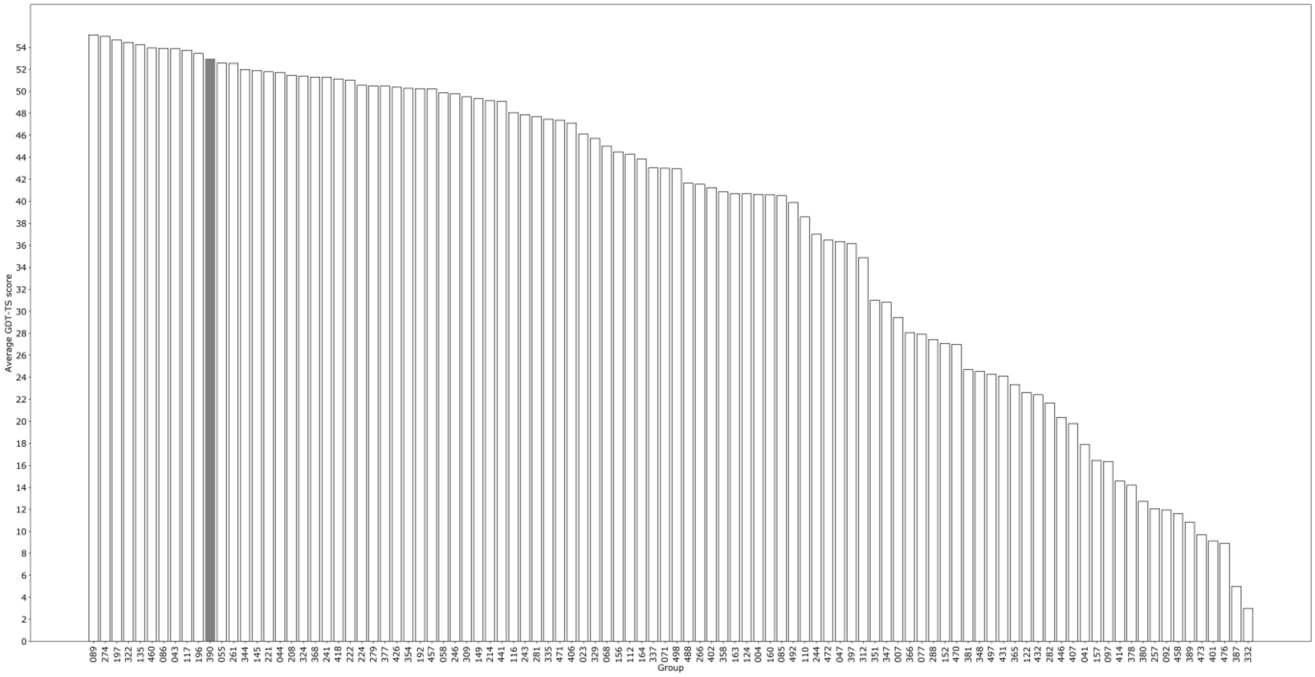


Figure 1. Performance of Bhattacharya human group (390) in the tertiary structure prediction category for 11 CASP13 "all groups" targets. The groups are sorted from left to right based on the average GDT-TS score of the first submitted model in non-increasing order. Bhattacharya human group is marked in gray.

Availability

scoreD, clustQ, and refineD methods are freely available at http://watson.cse.eng.auburn.edu/scoreD/, http://watson.cse.eng.auburn.edu/clustQ/, and http://watson.cse.eng.auburn.edu/refineD/ respectively.

1. Bhattacharya, D. & Shuvo, M. H. scoreD: Deep discriminative binary classifier ensemble for protein scoring. Submitted (2018).
2. Alapati, R. & Bhattacharya, D. in Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. 307-314 (ACM).
3. Bhattacharya, D. refineD: Improved protein structure refinement using machine learning based restrained relaxation. Submitted (2018).

clustQ: Multi-model QA using superposition-free weighted internal distance comparisons

Rahul Alapati, Md Hossain Shuvo, Debswapna Bhattacharya

Department of Computer Science and Software Engineering, Auburn University, Auburn, AL 36849, USA.

bhattacharyad@auburn.edu

We developed a new multi-model QA method, clustQ[1], by computing average pairwise similarity of a decoy with respect to the decoy pool using superposition-free weighted internal distance comparisons. clustQ was tested in CASP13 as "Bhattacharya-ClustQ".

Methods

We extended the Q-score[2] originally introduced by CASP8 assessors to propose a weighted version called WQ-score, based on weighted internal distance comparisons at four different sequence separations $Q_{narrow,}$ $Q_{short,}$ Q_{medium} and Q_{long}; obtained by averaging the Q_{ij} for each pair of residues i, j that satisfy $|i-j| < 6$, $6 \leq |i-j| < 12$, $12 \leq |i-j| < 24$ and $24 \leq |i-j|$ respectively. The weights were assigned as 1, 2, 4 and 8 for $Q_{narrow,}$ $Q_{short,}$ Q_{medium} and Q_{long} respectively. Higher weights were assigned to residues far away in the sequence because such long-range interactions carry more information about the overall protein fold than local short-range interactions. clustQ performed all against all pairwise comparisons of server models using WQ-score in order to estimate accuracy of decoy based on average WQ score.

Results

In Figure 1, we present per-target Pearson correlation and loss for Bhattacharya-ClustQ with respect to GDT-TS and GDT-HA for 12 targets that could be identified in PDB as of writing this abstract. It shows that Bhattacharya-ClustQ is well correlated with both GDT-TS and GDT-HA (average per-target correlation ~ 0.85). Loss is less than 0.1 GDT points for most targets.

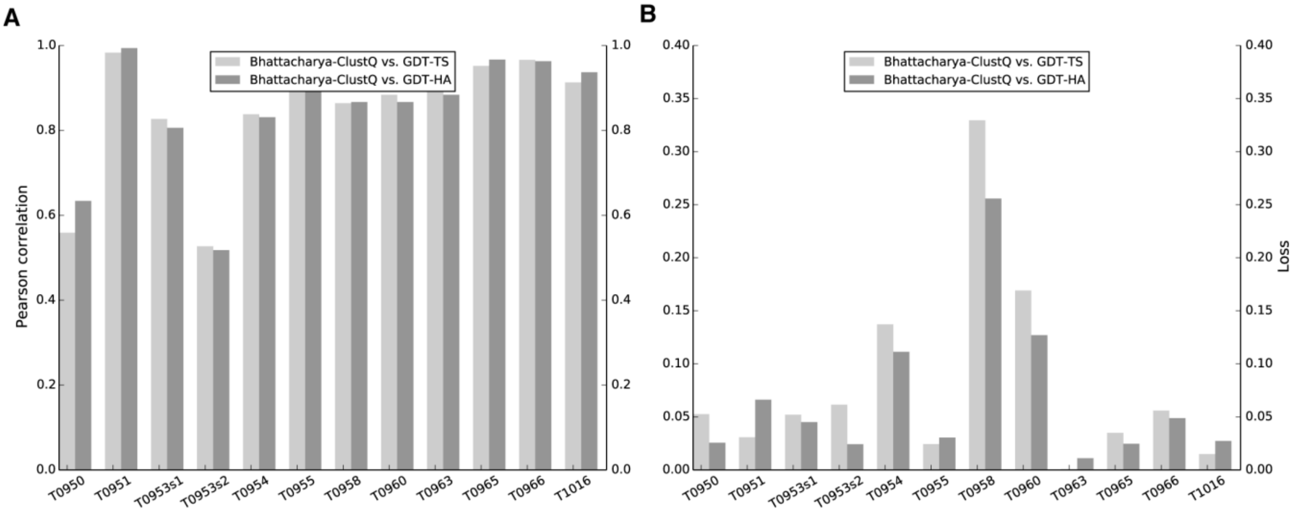


Figure 1. Performance of multi-model QA method Bhattacharya-ClustQ for 12 CASP13 targets. (A) Per-target Pearson correlation with respect to GDT-TS and GDT-HA, (B) GDT-TS and GDT-HA loss.

Availability

clustQ webserver and standalone version are freely available at http://watson.cse.eng.auburn.edu/clustQ/.

1. Alapati, R. & Bhattacharya, D. in Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. 307-314 (ACM).
2. Ben-David, M. et al. Assessment of CASP8 structure predictions for template free targets. Proteins: Structure, Function, and Bioinformatics 77, 50-65 (2009).

refineD: Protein structure refinement using machine learning guided restrained relaxation

Debswapna Bhattacharya, Md Hossain Shuvo

Department of Computer Science and Software Engineering, Auburn University, Auburn, AL 36849, USA.

bhattacharyad@auburn.edu

We developed a new protein structure refinement method, refineD[1], by predicting multi-resolution probabilistic restraints from the starting structure using our newly developed machine learning based binary classifier ensemble, scoreD[2], and subsequently converting these restraints into scoring term to guide conformational sampling during structure refinement. refineD was tested in CASP13 as "Bhattacharya-Server". The proposed predictor, for the first time, applies machine learning derived multi-resolution probabilistic restraints in protein structure refinement.

Methods

We used four restraint resolutions as adopted in GDT-HA (0.5, 1, 2, and 4Å), centered on the C_α atom of each residue that were predicted by ensemble of four deep discriminative classifiers trained using combinations of sequence and structure-derived features as well as several energy terms from Rosetta centroid scoring function[3]. Output from the ensemble of four classifiers were subsequently converted to multi-resolution probabilistic restraints and integrated as additional scoring term to Rosetta's all-atom energy function[4] to perform restrained relaxation using the FastRelax application of Rosetta[5,6].

Given a starting structure for refinement, each multi-resolution restraint was individually applied in a non-cumulative manner for all residues weighted according to their probabilities as predicted by the binary classifier ensemble. We employed restrained FastRelax protocol for five iterations using four parallel threads to generate a total of twenty refined models, each parallel thread using a different restraint resolution. We subsequently used probabilistic combination of the binary classifiers to select the top five high scoring models (refer Bhattacharya-Server QA abstract) to be submitted as refined structures.

Availability

refineD webserver is freely available at http://watson.cse.eng.auburn.edu/refineD/.

1. Bhattacharya, D. refineD: Improved protein structure refinement using machine learning based restrained relaxation. Submitted (2018).
2. Bhattacharya, D. & Shuvo, M. H. scoreD: Deep discriminative binary classifier ensemble for protein scoring. Submitted (2018).
3. Rohl, C. A., Strauss, C. E., Misura, K. M. & Baker, D. Protein structure prediction using Rosetta. Methods in enzymology 383, 66-93 (2004).
4. Alford, R. F. et al. The Rosetta all-atom energy function for macromolecular modeling and design. Journal of chemical theory and computation 13, 3031-3048 (2017).
5. Khatib, F. et al. Algorithm discovery by protein folding game players. Proceedings of the National Academy of Sciences 108, 18949-18953 (2011).
6. Leaver-Fay, A. et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods in enzymology 487, 545 (2011).

scoreD: Estimating Global Distance Test using deep discriminative binary classifier ensemble

Debswapna Bhattacharya, Md Hossain Shuvo

Department of Computer Science and Software Engineering, Auburn University, Auburn, AL 36849, USA.

bhattacharyad@auburn.edu

Global Distance Test (GDT)[1], one of the most widely used measures for computing accuracy of decoy, categorizes the alpha-carbon atom (C_α) of each residue of a decoy to be within a fixed number of predefined distance thresholds with respect to the native state after optimal structural superposition. We developed a new single-model QA method, scoreD[2], by modeling GDT-TS score using deep discriminative binary classifier ensemble. scoreD was tested in CASP13 as "Bhattacharya-SingQ". A variant of scoreD targeted at modeling GDT-HA score was also tested as "Bhattacharya-Server". The proposed predictors, for the first time, apply binary classification paradigm for modeling GDT score.

Methods

We used Deep Convolutional Neural Fields (DeepCNF)[3,4], a deep discriminative learning classifier, to predict the likelihood of C_α atom of any residue of a decoy to be within rÅ with respect to the native. In order to model GDT-TS score, we trained an ensemble of four DeepCNF binary classifiers after fixing r to 1, 2, 4, 8Å; and subsequently performed probabilistic weighted averaging to predict the overall accuracy score of a decoy (a.k.a. scoreD). For modeling GDT-HA score, we trained four different DeepCNF binary classifier ensemble after fixing r to 0.5, 1, 2, 4Å that were then probabilistically combined to estimate the overall accuracy score. Consequently, the proposed predictors are probabilistic equivalents of GDT measure.

Each DeepCNF classifier combined several centroid scoring functions of Rosetta[5], sequence profile based residue conservation features as well as the consistency measures between structural features extracted from the decoy conformation and predicted from the decoy's primary sequence; to be trained using datasets culled from 3DRobot[6] structural decoys. We specifically chose DeepCNF classifiers because our datasets suffered from class imbalance problem that is particularly pronounced at the highest and lowest distance thresholds and DeepCNF has been show to be particularly well suited for learning from imbalanced datasets by directly maximizing the empirical Area Under the ROC Curve (AUC), which is an unbiased measurement for imbalanced data[7].

Results

We use twofold evaluation criteria to quantitate the performance of single-model QA: (i) ability to reproduce the true decoy-native similarity scores, and (ii) ability to find the best decoy. For the first criterion, we use per-target Pearson correlation between all decoys' true GDT values and its estimated scores. Consequently, higher correlation indicates better performance. For the second criterion, we use average GDT loss that is the difference between the true accuracy score of the top decoy selected by a scoring function and that of the best possible decoy in the decoy pool in terms of GDT. A lower loss, therefore, indicates better performance.

In Figure 1, we present per-target Pearson correlation and loss for Bhattacharya-SingQ and Bhattacharya-Server with respect to GDT-TS and GDT-HA respectively for 12 targets that could be identified in PDB as of writing this abstract. It shows that Bhattacharya-SingQ (average per-target correlation with GDT-TS 0.65) has slightly better correlation than Bhattacharya-Server (average per-target correlation with GDT-HA 0.64). In terms of loss, except for few targets (e.g. T0955 and T0958) the performance is comparable.

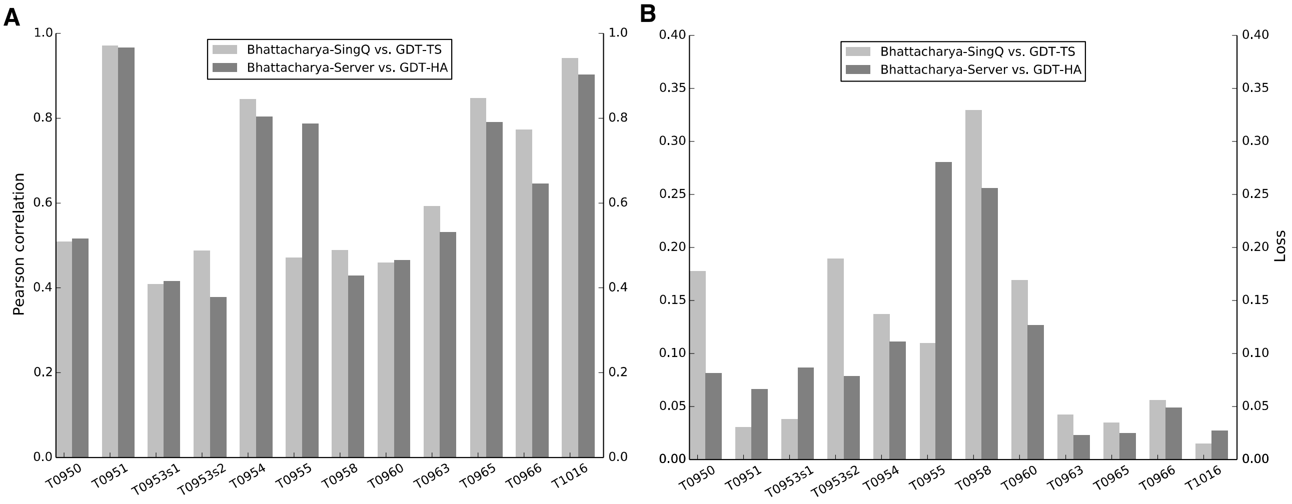


Figure 1. Performance of single-model QA methods Bhattacharya-SingQ and Bhattacharya-Server for 12 CASP13 targets. (A) Per-target Pearson correlation between Bhattacharya-SingQ vs. GDT-TS and Bhattacharya-Server vs. GDT-HA, (B) GDT-TS loss for Bhattacharya-SingQ and GDT-HA loss for Bhattacharya-Server.

Availability

scoreD webserver is freely available at http://watson.cse.eng.auburn.edu/scoreD/.

1. Zemla, A. LGA: a method for finding 3D similarities in protein structures. Nucleic acids research 31, 3370-3374 (2003).
2. Bhattacharya, D. & Shuvo, M. H. scoreD: Deep discriminative binary classifier ensemble for protein scoring. Submitted (2018).
3. Wang, S., Weng, S., Ma, J. & Tang, Q. DeepCNF-D: predicting protein order/disorder regions by weighted deep convolutional neural fields. International journal of molecular sciences 16, 17315-17330 (2015).
4. Wang, S., Peng, J., Ma, J. & Xu, J. Protein secondary structure prediction using deep convolutional neural fields. Scientific reports 6, 18962 (2016).
5. Rohl, C. A., Strauss, C. E., Misura, K. M. & Baker, D. Protein structure prediction using Rosetta. Methods in enzymology 383, 66-93 (2004).
6. Deng, H., Jia, Y. & Zhang, Y. 3DRobot: automated generation of diverse and well-packed protein structure decoys. Bioinformatics 32, 378-387 (2015).
7. Wang, S., Sun, S. & Xu, J. in Joint European Conference on Machine Learning and Knowledge Discovery in Databases. 1-16 (Springer).

Protein structure refinement using intermediate resolution, coarse-grained model, knowledge-based energy function(s), Monte Carlo methods, supported by MQAP constraints

M.J. Boniecki[1]

1 - International Institute of Molecular and Cell Biology in Warsaw, ks. Trojdena 4 street, 02-109 Warsaw, Poland.

mboni@genesilico.pl

In this CASP, I participated only in refinement category. I employed the module for protein structure prediction, which is part of a method for 3D modeling of protein-RNA complexes. The method was developed in prof. Bujnicki lab, we named it SimRNP[1].

Methods

SimRNP is recently developed method for modeling of proteins, RNAs, and protein-RNA complexes. SimRNP uses a coarse-grained representation of protein and RNA molecules, utilizes the Monte Carlo method to sample the conformational space, and relies on a statistical potential to describe the interactions in the folding process. It allows for modeling of complex formation for assemblies comprising two or multiple protein and RNA chains. Modeling system can be supported by various types of restraints, that can be derived from biological experiments or just restrains the limit of possible deformation of a given parts of the modeling system.

In the protein module, a protein backbone is represented by C-alpha atoms only, while side groups are represented by several pseudo-atoms, depending on the size of a side group. The energy function is a statistical potential. Protein backbone propensities are controlled by sequence dependent statistical energy therms, while side chains interactions are controlled by contact/distance dependent statistical potential. We also developed dedicated a set of Monte Carlo moves that are used by a conformation sampling engine.

In CASP13, I scored both global and local quality of input models using quality assessment method MQAP[2], I relied also on hints provided by the Organizers. I converted results of MQAP evaluation into constraints. I was running simulations using Replica Exchange Monte Carlo Method (REMC). Finally, I clustered results and score them using MQAP method.

SimRNP is developed on SimRNA framework[3].

Availability

SimRNP is still under development. It will be publicly available after publication.

1. Bujnicki,J., Boniecki,M., (2017), SimRNP: a new method for fully flexible modeling of protein-RNA complexes and for simulations of RNA-protein binding. *FEBS JOURNAL.* 284, SI Suppl.: 1, 201-201 Meeting Abstract: P.1.3-066.
2. 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.
3. Boniecki,M.J., Lach,G., Dawson,W.K., Tomala,K., Lukasz,P., Soltysinski,T., Rother,K.M., Bujnicki.J.M., (2016), SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. *Nucleic Acids Res*, **44**, e63.

Collaborative de novo protein structure prediction using stepwise fragment sampling with help of contact prediction and model selection based on deep learning techniques

Renzhi Cao[1*], Dong Si[2], Yajun An[3], Leong Chan[4], Ben Chen[4], Badri Adhikari[5], Xiaoyan Hao[6], Jumin Zhao[6], Nicholas Crossman[1], Emily Shane[1], Natalie Stephenson[1], Connor Whyte[1], Kyle Hippe[1], Rachel Schmit[1], Max Staples[1], John Smith[1], Matthew Conover[1], Yutong Fan[7], Davis Railsback[3], Yiheng Fu[8], Xi Chen[9], and Xiuqi Cao[10]

1 - Department of Computer Science, Pacific Lutheran University, Tacoma, WA 98447; 2 - Division of Computing and Software Systems, University of Washington-Bothell, Bothell, WA 98011; 3 - School of Interdisciplinary Arts and Sciences, University of Washington-Tacoma, Tacoma, WA 98402; 4 - School of Business, Pacific Lutheran University, Tacoma, WA 98447; 5 - Depart of Computer Science, University of Missouri-St. Louis, MO 63121; 6 - Department of Computer Engineering, Taiyuan University of Technology, Taiyuan, Shanxi 030600; 7 - Department of Social Sciences, Pacific Lutheran University, Tacoma, WA 98447; 8 - Department of Computing Science, University of Alberta, Edmonton, AB T6G 2R3; 9 - Department of software and information, University of Electronic Science and Technology of China 610054, China; 10 - College of Software, Nankai University, Tianjin 300457, China

*caora@plu.edu

In CASP 13, we blindly tested our new *de novo* protein structure prediction pipeline as a collaborative research, because most contributors in this research project are from primarily undergraduate institutions while computational resources are limited (our main server has only 24 CPUs). Instead of randomly sampling protein conformation space, stepwise fragment sampling is used in this method as it is more efficient and accurate[1,2]. Also, the contact information is incorporated in our pipeline, since contact prediction plays an important role in structure modeling in the recent CASP experiments[3–5]. Finally, deep learning technique is used for selecting 5 models as the final prediction of our method[6].

Methods

Step 1, contact prediction is made for each protein sequence. We used the latest version of MetaPSICOV2[3] to make contact prediction from the input protein sequence. We would like to mention that MetaPSICOV2 may fail occasionally, in this case, we use the alternative contact prediction from CCMpred and FreeContact[7,8].

Step 2, after the contact prediction was done, a request was sent to all connected computers for united-residue conformational search via stepwise and probabilistic sampling with the help of Unicon3D tool[1]. The secondary structure prediction and contact prediction from previous step was used in Unicon3D for *de novo* protein structure prediction, and all predictions were sent to main server before due date.

Step 3, compared to randomly sampling like Monte-Carlo search, sequential search turned to be more efficient and accurate. The main server did sequential protein conformational search with the help of SAINT2 tool[2]. The fragment used in this step was generated by modified version of FRAGSION tool[9], which is ultra-fast and accurate in fragment generation based on Hidden Markov Model. Because of computational resource limitation, we only generated fragment with size 8 and 12. The contact prediction from first step was also used to guide the protein structure prediction process.

Step 4, model selection from thousands of protein decoys is crucial in protein structure prediction. Qprob[10] is a super-fast tool to rank all decoys based on the model quality, and we select top 100 decoys based on Qprob's ranking. After that, we use deep learning technique (DeepQA tool[6]) with the help of clustering for diversity[11] to select 5 models as our final prediction.

Availability

The Cao-server is available at the following link:
https://www.cs.plu.edu/~caora//index.php/Cao_server/

1. Bhattacharya, D., Cao, R. & Cheng, J. (2016). UniCon3D: de novo protein structure prediction using united-residue conformational search via stepwise, probabilistic sampling. *Bioinformatics* **32**, 2791–2799
2. de Oliveira, S.H.P., Law, E.C., Shi, J. & Deane, C.M. (2018). Sequential search leads to faster, more efficient fragment-based de novo protein structure prediction. *Bioinformatics* **34**, 1132–1140
3. Buchan, D.W.A. & Jones, D.T. (2017). Improved protein contact predictions with the MetaPSICOV2 server in CASP12. *Proteins: Struct. Funct. Bioinf.* **86**, 78–83
4. Adhikari, B., Hou, J. & Cheng, J. (2018). Protein contact prediction by integrating deep multiple sequence alignments, coevolution and machine learning. *Proteins* **86 Suppl 1**, 84–96
5. Schaarschmidt, J., Monastyrskyy, B., Kryshtafovych, A. & Bonvin, A.M.J.J. (2018). Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age. *Proteins* **86 Suppl 1**, 51–66
6. Cao, R., Bhattacharya, D., Hou, J. & Cheng, J. (2016). DeepQA: improving the estimation of single protein model quality with deep belief networks. *BMC Bioinformatics* **17**, 495
7. Seemayer, S., Gruber, M. & Söding, J. (2014). CCMpred--fast and precise prediction of protein residue-residue contacts from correlated mutations. *Bioinformatics* **30**, 3128–3130
8. Kaján, L., Hopf, T.A., Kalaš, M., Marks, D.S. & Rost, B. (2014). FreeContact: fast and free software for protein contact prediction from residue co-evolution. *BMC Bioinformatics* **15**, 85
9. Bhattacharya, D., Adhikari, B., Li, J. & Cheng, J. (2016). FRAGSION: ultra-fast protein fragment library generation by IOHMM sampling. *Bioinformatics* **32**, 2059–2061
10. Cao, R. & Cheng, J. (2016). Protein single-model quality assessment by feature-based probability density functions. *Sci. Rep.* **6**, 23990
11. Cao, R., Bhattacharya, D., Adhikari, B., Li, J. & Cheng, J. (2015). Large-scale model quality assessment for improving protein tertiary structure prediction. *Bioinformatics* **31**, i116–i123

Combining evolution- and geometry-driven interface predictions with knowledge-based distance-dependent potentials for template-free modeling in CAPRI round 46

E. Laine[1], A. Carbone[1,2], and S. Grudinin[3]

1 - Sorbonne Université, CNRS, IBPS, Laboratoire de Biologie Computationnelle et Quantitative (LCQB), 75005 Paris, France, 2 - Institut Universitaire de France (IUF), Paris, France, 3 - Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LJK, 38000 Grenoble, France.*

elodie.laine@upmc.fr, sergei.grudinin@inria.fr

Protein-protein interfaces display specific evolutionary, physico-chemical and/or geometrical properties. We have previously developed JET^2, a method exploiting these properties to predict different types of protein-protein interfaces with high precision[1]. Given a query protein, JET^2 performs a sequence- and structure-based analysis of its surface to identify residues either conserved through evolution, often found in experimental interfaces, protruding to the solvent, or displaying a combination of some of these properties. It then clusters these residues based on 3D proximity to define patches likely to be participating in an interaction. JET^2 does not use any information coming from the potential partners of the query protein, it does not look for coevolution signals nor predict inter-protein contacts, contrary to Direct Coupling Analysis[2] (DCA)-like methods. The algorithm for computing residue conservation levels relies on a discrete combinatorial paradigm to randomly sample small subsets of sequences and explicitly accounts for the topology of the distance trees relating these sequences[3]. Hence, our conservation measure is markedly different from more popular statistical measures, *e.g.* those relying on information entropy. It has the advantage of being able to capture signals even on rather small sets of closely related sequences. JET^2 was applied to more than 20 000 protein chains (predictions available at: http://www.jet2viewer.upmc.fr/) and achieved 76% accuracy on more than 15 000 experimentally determined protein interfaces[4]. Here, we combined JET^2 predictions with the knowledge-base distance-dependent potentials KSENIA[5] and SBROD[6] for the prediction and scoring stages of the CASP13-CAPRI experiment.

Methods

JET^2 implements three scoring strategies aimed at detecting different types of protein-protein interfaces. Each strategy combines in a very straightforward way three sequence and structure-based residue descriptors to define one or several protein surface patches. A confidence score is assigned to each residue within each patch. Patches generated by different strategies may be included in one another, partially overlapping or distinct. Such a description of protein surfaces is particularly useful for the prediction of large complexes, where each chain interacts with several partners via distinct regions displaying different properties. For example, JET^2 is able to detect a homodimeric interface and an enzyme-substrate binding site on the same protein surface and distinguish them [1].

Results

In the prediction stage of the CASP13-CAPRI experiment, we used Hex[7] and SAM[8] rigid-body fast Fourier transform-accelerated docking engines to generate a vast amount of putative binding poses. We used the 50 best stage-2 server predictions, as ranked by the SBROD model quality assessment function, as starting docking models, and performed 1275 cross-docking runs for heterooligomers and 50 runs for homooligomers. For heterooligomeric assemblies we used Hex, for homooligomeric assemblies we used SAM, and for the mixed stoichiometries we used a combination of two. The obtained complexes were optimized using the KSENIA potential. In parallel of the docking procedure, JET^2 was run on the starting docking models.

Then, we combined JET^2 predictions and SBROD scores to rank and select docking conformations either generated by us (prediction stage) or by all the participants (scoring stage). For the scoring stage, we mapped the previously computed JET^2 predictions onto the provided docking conformations. The compliance of a given docking conformation with JET^2 predictions was evaluated by summing up JET^2 confidence scores over the ensemble of residues lying in the docked interface. By default, we considered all predictions generated by the three JET^2 scoring schemes and averaged the obtained by-residue confidence scores. For some targets, especially large assemblies, we manually chose one or several scoring schemes and combined them in an *ad-hoc* fashion, based on literature search. The final score was a linear combination of the normalized JET^2 score and the normalized SBROD score, such that each of the two scores contributes equally to the final result.

Availability

JET^2 is available at: http://www.lcqb.upmc.fr/jet2/JET2.html.

1. Laine E., Carbone A. (2015) Local geometry and evolutionary conservation of protein surfaces reveal the multiple recognition patches in protein-protein interactions. *PLoS Comput. Biol.* **11**:e1004580.
2. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. Identification of direct residue contacts in protein–protein interaction by message passing. Proc Natl Acad Sci USA. 2009;106:67–72.
3. Engelen S., Trojan L.A., Sacquin-Mora S., Lavery R., Carbone A. (2009) Joint evolutionary trees: a large-scale method to predict protein interfaces based on sequence sampling. *PLoS Comput. Biol.* **5**:e1000267.
4. Ripoche H., Laine E., Ceres N., Carbone A. (2017) JET2 Viewer: a database of predicted multiple, possibly overlapping, protein-protein interaction sites for PDB structures. *Nucleic Acids Research*. **45(D1)**:D236-D242.
5. Popov P., Grudinin, S (2015) Knowledge of Native Protein–Protein Interfaces Is Sufficient To Construct Predictive Models for the Selection of Binding Candidates. *J. Chem. Inf. Model*, **55:** 2242-2255.
6. Karasikov M., Pages G., Grudinin, S (2018) Smooth orientation-dependent scoring function for coarse-grained protein quality assessment. Unpublished.
7. Ritchie D.W., Kemp, G.J.L. (2000) Protein Docking Using Spherical Polar Fourier Correlations. *Proteins: Struct. Funct. Genet.* **39**: 178-194.
8. Ritchie D.W., Grudinin, S (2016) Spherical polar Fourier assembly of protein complexes with arbitrary point group symmetry. *J. Appl. Cryst*. **12** :1019-1028.

Hybrid ClusPro server in 2018 CASP/CAPRI rounds

Dima Kozakov[1], Sandor Vajda[2,3], Kathryn Porter[2], Dzmitry Padhorny[1], Israel Desta[2], Dmitri Beglov[2], Mikhail Ignatov[1], Sergey Kotelnikov[1,4]

1-Laufer Center for Physical and Quantitative Biology, Stony Brook University; Departments of 2-Biomedical Engineering and 3-Chemistry, Boston University; 4-Moscow Institute of Physics and Technology

The original ClusPro server performs rigid body docking using the PIPER program and clusters the 1000 lowest energy structures. The models are ranked according to cluster size. In order to deliver results to the user within 24 hours of submission, the current implementation of ClusPro does not include refinement beyond minimizing the energy of structures to remove steric overlaps. In spite of this limitation, the server has almost 7800 registered users, and run about 200,000 jobs in the last 3 years. In the recent years we have enhanced ClusPro with capabilities of accounting for additional information to restrain the search, including SAXS data and XL-MS cross-links.

In the latest rounds of the CASP-CAPRI experiment we have expanded the ClusPro server to use template based information when available. Based on the target sequence we identify structures that can serve as templates for the complex, and perform homology modeling based on the biological units of the templates. If no template is available, we perform free docking as described above. The server has the option of accepting pre-selected templates as input. In addition, we explore the option of further refining and validating template based models with free docking.

Methods

Model preparation.

Based on the sequence of the target we automatically detect available templates using HHPred, and identify those that contain homologs of the interacting biological unit to be predicted. If no template of the complex is found, we suggest to perform free docking. Since free docking by ClusPro requires three-dimensional structures as the input, we either use the HHPRED top template or in difficult cases build a "consensus" model for each target using the 150 server models provided by the CASP management committee. For each "easy" target most models had the same fold, with variations in loops and tails. Removal of the uncertain regions resulted in reliable "consensus" models that were used for docking.

Template based docking.

If a template of the biological complex is found then we model each monomer of the complex using Modeller, align separately to the template and co-minimize the resulting complex. Per rules of CAPRI we generate up to 10 models.

Free Docking.

Our free docking approach consists of two steps. The first step is running PIPER, a docking program that performs systematic search of complex conformations on a grid using the fast Fourier transform (FFT) correlation approach. The scoring function includes van der Waals interaction energy, an electrostatic energy term, and desolvation contributions calculated by a pairwise potential.

The second step of the algorithm is clustering the top 1000 structures generated by PIPER using pairwise RMSD as the distance measure. The radius used in clustering is defined in terms of C_α interface RMSD. For each docked conformation we select the residues of the ligand that have any atom within 10 Å of any receptor atom, and calculate the C_α RMSD for these residues from the same residues in all other 999 ligands. Thus, clustering 1000 docked conformations involves computing a 1000 × 1000 matrix of pairwise C_α RMSD values. Based on the number of structures that a ligand has within a (default) cluster radius of 9 Å RMSD, we select the largest cluster and rank its cluster center as number one. The members of this cluster are removed from the matrix, and we select the next largest cluster and rank its center as number two, and so on. After clustering with this hierarchical approach, the ranked complexes are subjected to a straightforward (300 step and fixed backbone) van der Waals minimization using the CHARMM potential to remove potential side chain clashes. ClusPro outputs the centers of the 10 largest clusters, which were submitted as predictions.

TopContact: Protein contact meta-predictions through machine learning

S. Schott-Verdugo, D. Mulnaes, H. Gohlke

CPCLab, Institut für Pharmazeutische und Medizinische Chemie, Heinrich-Heine-Universität Düsseldorf, Düsseldorf, Germany

gohlke@hhu.de

The incredible speed at which the number of sequenced proteomes increases led to the majority of the available protein sequences not having a proper template for making accurate homology models. This development explains why *ab initio* protein structural modelling currently has such an important role. In the last CASP competition, it was shown how the incorporation of machine learning methods to obtain residue-residue protein contacts increased the accuracy of the predictions in a remarkable manner[1]. Interestingly, the results and performances between multiple softwares are not necessarily overlapping, and the number of contacts required to correctly fold a protein are highly dependent on the target or the software[1].

Methods

To take advantage of the non-overlapping results of available contact prediction softwares, of predictors for protein structural characteristics, and of current machine learning algorithms, here, we present TopContact, a machine learning meta-predictor based at present on twelve state-of-the-art contact predictors, and four secondary structure and solvent accessibility predictors. The method was trained using a multi-staged Resnet-152 deep residual convolutional neural network architechture[2], using each contact centered in an 11x11 image, with the primary predictors and features as 127 image channels. For the training, an exhaustive dataset composed of 3237 randomly chosen proteins was culled, making a total of ~80 million residue pairs.

In a first stage, all residue pairs that are confidently identified as true negatives are discarded, leaving the remaining ones to be further filtered in a second stage, including the results from the first stage as a primary predictor. The predictions obtained from the second stage are filtered according to a sequence length cutoff, and the number of optimal contacts are defined on a per-protein basis. An automated workflow was established to apply the method, including an automated folding protocol with CONFOLD[3]. An additional stage that considers protein structural consistency is being included.

Results

Preliminary results on the training set show a higher average F1 score compared to all the contact predictors used, irrespective of the selected cutoff for the top ranked contacts. Additionally, a higher Precision-Recall AUC over all predicted contacts shows that TopContact works as a better classifier for true contacts than any of the primary predictors.

Availability

The method will be made available in early 2019 as a webserver at http://cpclab.uni-duesseldorf.de/. The preliminary method can be applied upon request by sending an e-mail to schottve@hhu.de.

1. Schaarschmidt, J.; Monastyrskyy, B.; Kryshtafovych, A.; Bonvin, A., Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age. *Proteins* 2018, **86** Suppl 1, 51-66.
2. He, K.; Zhang, X.; Ren, S.; Sun, J., Deep Residual Learning for Image Recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE: 2016; Vol. 77, pp 770-778.
3. Adhikari, B.; Bhattacharya, D.; Cao, R.; Cheng, J., CONFOLD: Residue-residue contact-guided ab initio protein folding. *Proteins: Structure, Function, and Bioinformatics* 2015, **83**, 1436-1449.

CPClab (TS)

TopScore: Using deep neural networks and large diverse datasets for accurate protein model quality assessment

D. Mulnaes[1] and H. Gohlke[1,2]

1- Institute for Pharmaceutical and Medicinal Chemistry, Department of Mathematics and Natural Sciences, Heinrich Heine University Düsseldorf, Düsseldorf, Germany, 2- John von Neumann Institute for Computing (NIC), Jülich Supercomputing Centre (JSC) & Institute for Complex Systems - Structural Biochemistry, Forschungszentrum Jülich GmbH, Jülich, Germany

daniel.mulnaes@hhu.de

The value of protein models obtained with automated protein structure prediction depends primarily on their accuracy. Protein model quality assessment is thus critical to select the model that can best answer biologically relevant questions from an ensemble of predictions. However, despite many advances in the field, different methods capture different types of errors, begging the question of which method to use. We introduce TopScore, a meta Model Quality Assessment Program (meta-MQAP) that uses deep neural networks to combine scores from fifteen different primary predictors to predict accurate residue-wise and whole-protein error estimates. TopScore competed in CASP13 and provided predictions for 87 targets.

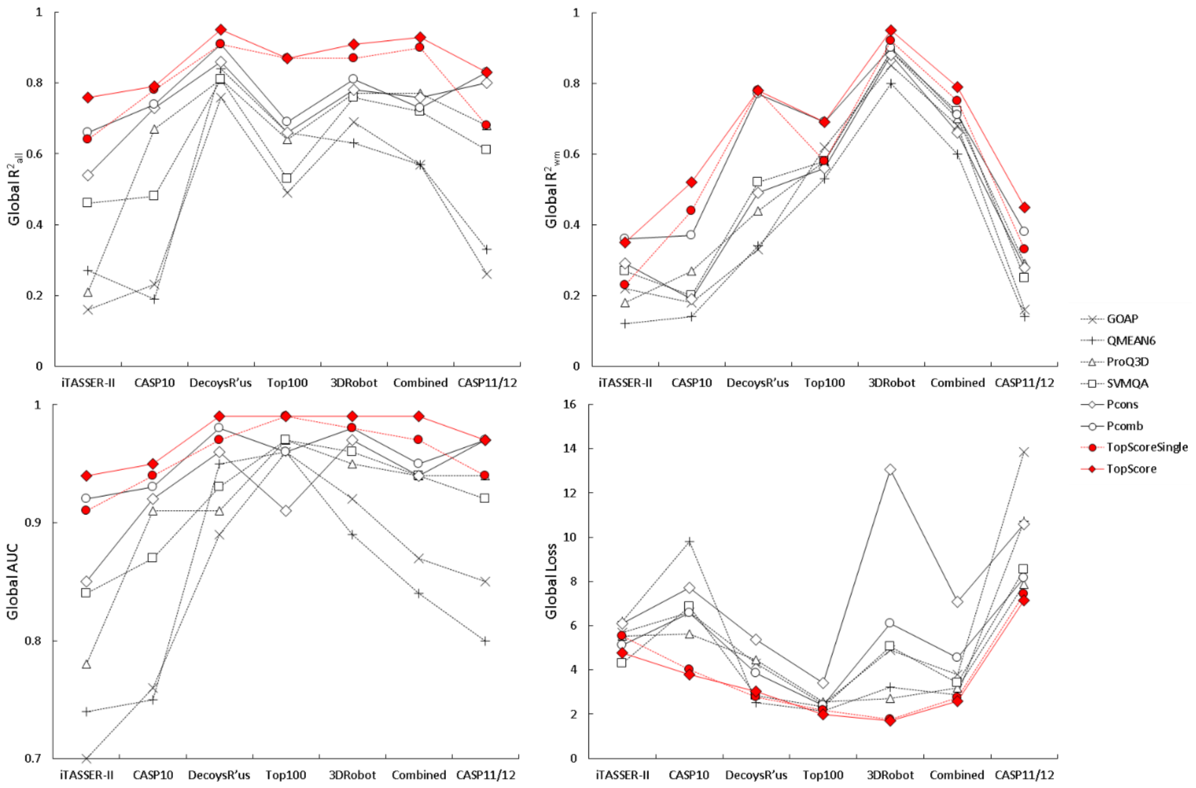


Figure 1. TopScore global performance. TopScore (red circles) and TopScoreSingle (red dashes) global performance compared to a subset of primary predictors (black). Dashed lines represent single-model methods and full lines methods that use clustering information. The 95% confidence intervals were calculated using the Fischer r-to-z transformation. The widest confidence interval for any R_{all}^2 or R_{wm}^2 was 0.01 and 0.12, respectively. Statistical significance was determined by the two-sided Steiger test. Accordingly, the R_{all}^2 and R_{wm}^2 of TopScore and TopScoreSingle are significantly different from any primary MQAP for the combined dataset ($p < 0.05$). In terms of R_{all}^2, for the CASP11/12 dataset, TopScoreSingle is not significantly different from ProQ3D, and neither is TopScore when compared to Pcomb. See Table S4 and S5 for numerical values of all investigated MQAPs. See Table S3 for statistics of lDDT distributions of individual datasets.

Methods

TopScore uses deep neural networks to predict the whole-protein (global) and residue-wise (local) model quality of the input models. The primary predictors of TopScore include methods that evaluate protein stereochemistry, packing and clashes (PROCHECK[1] and MolProbity[2]) knowledge-based distance, contact and angle potentials (ANOLEA[3], ProSA2003[4], DOPE[5], and GOAP[6]), composite scoring functions (QMEAN6[8] and SELECTpro[9]) advanced machine learning methods (ProQ2[10, 11], ProQ2D[12], ProQ3D[12], and SVMQA[7]) and clustering methods (ModFOLDClust2[13], SPICKER[14] and Pcons[10]). Each primary predictor is first normalized using a deep neural network with only that predictors score as an input and 1-lDDT score as a target. Subsequently all normalized values are then used as input for a deep neural network to produce the final prediction. Two methods were trained, with and without clustering methods, termed TopScore and TopScoreSingle.

Results

The predictions on six large independent datasets are highly correlated to superposition-independent errors in the model, achieving a Pearson's R_{all}^2 of 0.93 and 0.78 for whole-protein and residue-wise error predictions, respectively. This is a significant improvement over any of the investigated primary MQAPs, demonstrating that much can be gained by optimally combining different methods and using different and very large datasets.

Availability

TopScore and TopScoreSingle are available from the authors upon request.

1. Laskowski, R.A., et al., PROCHECK: a program to check the stereochemical quality of protein structures. Journal of applied crystallography, 1993. 26(2): p. 283-291.
2. 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.
3. Melo, F. and E. Feytmans, Novel knowledge-based mean force potential at atomic level. Journal of molecular biology, 1997. 267(1): p. 207-222.
4. Sippl, M.J., Recognition of errors in three-dimensional structures of proteins. Proteins: Structure, Function, and Genetics, 1993. 17(4): p. 355-362.
5. Shen, M.y. and A. Sali, Statistical potential for assessment and prediction of protein structures. Protein science, 2006. 15(11): p. 2507-2524.
6. Zhou, H. and J. Skolnick, GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. Biophysical journal, 2011. 101(8): p. 2043-2052.
7. Manavalan, B. and J. Lee, SVMQA: support–vector-machine-based protein single-model quality assessment. Bioinformatics, 2017. 33(16): p. 2496-2503.
8. Benkert, P., M. Biasini, and T. Schwede, Toward the estimation of the absolute quality of individual protein structure models. Bioinformatics, 2011. 27(3): p. 343-350.
9. Randall, A. and P. Baldi, SELECTpro: effective protein model selection using a structure-based energy function resistant to BLUNDERs. BMC structural biology, 2008. 8(1): p. 52.
10. Wallner, B. and A. Elofsson, Identification of correct regions in protein models using structural, alignment, and consensus information. Protein Science, 2006. 15(4): p. 900-913.
11. Wallner, B. and A. Elofsson, Can correct protein models be identified? Protein science, 2003. 12(5): p. 1073-1086.
12. Uziela, K., et al., ProQ3D: improved model quality assessments using deep learning. Bioinformatics, 2017. 33(10): p. 1578-1580.
13. McGuffin, L.J. and D.B. Roche, Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments. Bioinformatics, 2010. 26(2): p. 182-188.
14. Zhang, Y. and J. Skolnick, SPICKER: A clustering approach to identify near-native protein folds. Journal of computational chemistry, 2004. 25(6): p. 865-871.

Template-based Protein Structure Prediction based on Profile-Profile Alignments

T. Nakamura[1], H. Fukuda[1], Y. Yamamori[2], Y. Tsuchiya[2] and K. Tomii[1,2]

1 -Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo , 2 - Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST)

k-tomii@aist.go.jp

Profile-profile comparison is a powerful method for template-based modeling because it not only can detect distantly related proteins but also can make alignments more accurate. We construct and evaluate 3D-models based on profile-profile alignments for a given sequence through our pipeline.

Methods

Our pipeline to construct and evaluate 3D-models for a target sequence contains five steps: 1) prediction of intrinsically disordered region (IDR) of the target protein by DISOPRED3[2], 2) profile construction for both the target sequence except IDR and template sequences, 3) profile-profile alignment and scoring calculated by FORTE[1] series, that are our own profile-profile comparison methods, 4) 3D-model construction based on the alignments using MODELLER[3], 5) evaluation of 3D-models using Verify3D[4] and dDFire[5].

To construct profiles of both targets and templates, we used PSI-BLASTexB[6], DELTA-BLAST[7] and HHblits[8]. PSI-BLASTexB is the revised version of PSI-BLAST[9] to obtain better position- specific scoring matrix (PSSM), as the original PSI-BLAST could produce inappropriate scores in PSSM derived from a narrow block. When we construct profiles with PSI-BLASTexB, we used three types of queries as follows: 1) as an input multiple sequence alignment (MSA) for PSI-BLASTexB, we made a MSA by using MPI-parallelized MAFFT[10,11] with homologous sequences detected by SSearch with MIQS[12] against NCBI nr database; 2) as an input MSA, we prepared it with structurally similar domains derived from SCOP/PDP domain definition by stacking pairwise alignments produced by TM-align[13]; 3) we made a profile using only a target/template sequence as an input for PSI-BLASTexB.

In our model selection step, first, 3D-models with very low scores (<20) of Verify3D were removed. Then, models with high Z-scores (>8) were picked up for further selection. For further selection, we used scores of Z-score, Verify3D, dDFire, and similarity with predicted secondary structure by RaptorX-Property[14]. Clustering based on pairwise TM-score was also done for models with high Z-scores. If there were no models with Z-score of more than 8, models with low Z-scores were also used for the selection. For multimeric targets, the stoichiometry of the template protein was considered to select a model. When there was no template that satisfies the stoichiometry of the target protein, rigid-body docking was performed using ZDOCK[15] or M-ZDOCK[16] to obtain multimeric form.

For the refinement targets of Small-angle X-ray scattering (SAXS), to rank and select 3D-models, we used the metrics between the calculated SAXS profile of a model and the provided experimental SAXS profile as well as other criteria mentioned above. As the metrics, we calculated Chi^2 and volatility of ratio (V_R) using FoXs[17]. In the cases of SAXS-assisted problem, the metrics between SAXS profiles are prefered to other criteria, and if Chi^2 was not consistent with V_R, models were selected based on V_R.

Results

For target T0955 which would be categorized to the template-based problem, we could construct and select the model with high GDT_TS (0.8354). For T0965, even we could construct a model with relatively high GDT_TS, we failed to select the one. For T1009, we succeeded in selecting the model with (near) highest GDT_TS in our models.

1. Tomii, K., Hirokawa, T. & Motono, C., (2005). Protein structure prediction using a variety of profile libraries and 3D verification. *Proteins*, **61 Suppl 7**, 114-21.
2. Jones, D. T., & Cozzetto, D., DISOPRED3: precise disordered region predictions with annotated protein-binding activity. (2015). *Bioinfomatics*, **31**, 857–863.
3. Webb, B. & Sali, A., (2014). Comparative Protein Structure Modeling Using MODELLER. *Curr Protoc Bioinformatics,* **47***,* 1-32.
4. Eisenberg, D., Luthy, R., Bowie, J. U., (1997). VERIFY3D: assessment of protein models with three-dimensional profiles. *Methods Enzymes*, **277**, 396-404.
5. Yang, Y. & Zhou, Y., (2008). Specific interactions for ab initio folding of protein terminal regions with secondary structures. *Proteins*, **72**, 793–803.
6. Oda, T., Lim, K., & Tomii, K., (2017). Simple adjustment of the sequence weight algorithm remarkably enhances PSI-BLAST performance. *BMC Bioinfomatics*, **18**, 288
7. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389-402.
8. 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.
9. Boratyn, G. M., Schaffer, A. A., Agarwala, R., Altschul, S. F., Lipman, D. J. & Madden, T. L. (2012). Domain enhanced lookup time accelerated BLAST. *Biol Direct*, **7**, 12.
10. Katoh, K. & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*, **30**, 772-80.
11. Nakamura, T., Yamada, K., Tomii, K., & Katoh, K., Parallelization of MAFFT for large-scale multiple sequence alignments. (2018), *Bioinformatics*, **34**, 2490–2492.
12. Yamada, K. & Tomii, K., (2014). Revisiting amino acid substitution matrices for identifying distantly related proteins. *Bioinformatics*, **30**, 317-25.
13. Zhang, Y. & Skolnick, J., (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res*, **33**, 2302-9.
14. Wang, S., Li, W., Liu, S., & Xu, J. (2016). RaptorX-Property: a web server for protein structure property prediction. *Nucleic Acids Research*, **44**, 430-5.
15. Chen, R., Li L., & Weng, Z., (2003) ZDOCK: An Initial-stage Protein Docking Algorithm. *Proteins*, **52** 80-7.
16. Pierce, B., Tong W., & Weng, Z., (2005) M-ZDOCK: A Grid-based Approach for C_n Symmetric Multimer Docking. *Bioinformatics*, **21**, 1472-1476.
17. Schneidman-Duhovny, D., Hammel, M., Tainer, J. A. & Sali, A. (2013). Accurate SAXS profile computation and its assessment by contrast variation experiments. *Biophysical journal*, **105**, 962–974.

MoDyCHoPS - Molecular Dynamics Coupled Homologue Protein Structure Refinement Protocol

Dennis Della Corte

Department of Physics and Astrophysics, Brigham Young University, Provo, UT, USA

dennis.della_corte@fulbrightmail.org

During CASP13, we applied a molecular dynamics (MD) protocol for the refinement of protein structures. In previous iterations of CASP9 and CASP11 the group of Gunnar Schröder (schroderlab) showed that coupling of structural homologs during MD can lead to improved refinement[1]. The method used by schroderlab in CASP11 [2] was adapted and further improved for CASP13. The major drawbacks of the original method were missing equilibrations, insufficient simulation lengths and poor stereochemistry in the final submissions. For each refinement target of CASP13 a PSIBLAST search was conducted to identify sequence homologues. The start structure was coupled through Calpha distance restraints to seven homologue structures build from the selected homologues sequences with MODELER. To prepare the system for production runs, two short equilibrations with constant volume and pressure were simulated. Following equilibration 20ns production MD simulations with solvation were ran for each target with GROMACS. This was repeated five time and the resulting trajectories were averaged to obtain five models. The quality of these models was further improved through a short MD with Ramachandran constraints in CNS, resulting in better MOLPROBITY [3] scores. The same CNS post processing was also applied to the starting structure. The resulting six structures were ranked according to their MOLPROBITY score with the highest score resulting in submission model 1. Only the lowest scoring model was not selected for submission.

The protocol was completely automated and generated five submissions for 27 of 29 refinement targets. Only for target R0949 human intervention was necessary due to an unexpected gap in the protein structure. Here we manually modeled the missing residues in a preparatory step before starting the simulation protocol. R0949 was the first oligomeric target seen at CASP refinement. Our tools were not designed to deal with multiple proteins in the start structure and time was insufficient to adapt the routines accordingly. For this target we only applied the post processing script to the start structure and submitted one single model.

1. Wildberg, A, Della Corte D., and Schröder G.F.. Coupling an ensemble of homologs improves refinement of protein homology models J. Chem. Theo. Comput. (2015), 11(12):5578-5582
2. Della Corte D., Wildberg A., and Schröder G.F.. Protein Structure Refinement with Adaptively Restrained Homologous Replicas Proteins (2015), doi:10.1002/prot.24939
3. Chen et al. (2010) MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallographica D66:12-21.

Combining orthodox sequence homology analysis with spectral analysis for protein 3D structure prediction

Carlos A. Del Carpio Muñoz

Graduate School of Medical Sciences. Doctoral Program in Biodefense. Nagoya City University, 1, Kawasumi, Mizuho-cho, Mizuho-ku, Nagoya 467-8601, Japan

delcmca@gmail.com

The widely accepted structural biological notion that protein structure conservation through natural evolution outpaces sequence conservation poses restrictions on the potentiality of methodologies oriented to predict the 3D structure of proteins based on sequence homology analysis alone. This assertion is particularly evident in the so-called twilight zone of sequence homology (20~30% of similarity), where prediction of protein 3D structure based only on sequence homology methodologies are frequently of limited success.

3D structures for CASP13 targets were predicted using a new protocol proposed by the author that combines orthodox homology methods with an improved technique that identifies folding patterns of proteins based on a spectral analysis of protein amino acid sequences. Partial analysis of the results shows a significant improvement in the prediction process, namely in medium difficulty targets.

Methodology

The author so far has proposed an original methodology to gauge for protein 3D structural similarities at the heart of which is an spectral representation of the sequences of amino acids represented quantitatively by the values of their different physicochemical properties[1; 2]. This has led to an automatic codification of folding patterns that can be used to retrieve patterns in a classification tree like the SCOP[3] groups and families of proteins. The methodology recognizes protein folding patterns comparing the encoded target protein sequence with the data base of SCOP families of protein sequences previously encoded following the proposed spectral protocol. This process plays a pivotal role in homolog identification for sequences of low similarity.

In CASP13, we have combined this methodology with orthodox sequence based homology as well as information obtained by secondary structure prediction methods. This has led to an improvement in the identification of folding patterns that were difficult to find employing any single methodology at a time. Prediction of the secondary structure assists in the assignation of the right sequence to any 3D piece of structure, namely when the selection of the protein is made by the spectral technique proposed by the author.

Here we discuss the effectiveness of our combined methodology when it is blindly applied to predict the 3D structures of CASP13 targets.

On the other hand, protein assemblies were also dealt using a new methodology developed by our group in recent years together with the system for protein-protein interaction assessment MIAX[4; 5].

Results

A remarkable improvement in the assignation of protein folding patterns can be observed for the CASP13 targets whose PDB structure have been released. Nevertheless a whole assessment of the proposed technique may require a larger set of targets with experimental structures.

1. Del Carpio, C. A. & Carbajal, J. C. (2002). Folding pattern recognition in proteins using spectral analysis methods. *Genome Inform* **13**, 163-72.
2. Del Carpio, C. A. & Yoshimori, A. (2002). *Fully automated protein tertiary structure prediction using Fourier*

transform spectral methods. Protein Structure Prediction: Bioinformatics, University of California, International University Line.

3. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* **247**, 536-40.
4. Del Carpio, C. A., Ichiishi, E., Yoshimori, A. & Yoshikawa, T. (2002). A new paradigm for modeling biomacromolecular interactions and complex formation in condensed pahses. *Proteins: Structure, Function, and Genetics* **48**, 696-732.
5. Del Carpio, C.A., Ichiishi E. (2017). Inference of Protein Multimeric Complex Dynamic Order of Formation: An Active Region Recognition Based Approach. *Intern. Journal of Genomics and Data Mining* **2017**, 1.

A deep-learning approach to protein structure prediction

Mu Gao, Hongyi Zhou and Jeffrey Skolnick

Center for the Study of Systems Biology, Georgia Institute of Technology

skolnick@gatech.edu

To predict the corresponding structural fold from the protein sequence is one of most challenging problems in computational biology. In this CASP, we introduce a novel computational approach, DESTINI, that combines a deep-learning algorithm for protein residue/residue contact prediction and template-based structural modeling. The application of deep-learning neural networks to protein contact prediction is an emerging, promising idea. However, it has not been taken advantage of by template-based structural modeling as described here.

Methods

DESTINI has two main components: contact prediction and structural modeling. The contact prediction is an implementation of a fully convolutional residual neural network composed of 102 layers in total, including 40 convolutional layers. The input features consist of three 2D features: co-evolutionary coupling scores[1], a statistical potential[2], mutual information for pairs of residues[3], and three 1D features: BLAST sequence profiles[4], secondary structure and solvent accessibility predictions[5], which are converted into 2D features by concatenating 1D features of two separate residues of each residue pair. The contact predictions are then supplied to structural modeling, the second component of DESTINI, which is a further development based on the TASSER^{VMT} approach[6]. When there is no suitable template model available, the structural modeling essentially makes *de novo* predictions[7]; if there is a significant structural template hit, modeling based on the template(s) is conducted. In both scenarios, confident contact predictions serve as the main driver towards the native structural fold. For the CASP competition, human intervention was applied to multiple domain targets, which was partitioned into individual domains according to the contact prediction of the full sequence and template threading results. Each domain was then modeled separately.

Results

In a large benchmark test on 606 “glass-ceiling” targets that are difficult for template-based approach as described previously[8], only considering the top1 model, DESTINI is capable of predicting native-like folds for 37% of targets, compared to only 9% of targets by TASSER. Among these targets, the mean TM-score is 0.539, indicating a highly likely correct fold, versus 0.456 by TASSER. Moreover, even for “easy” targets whose correct template is most likely revealed by threading algorithms, DESTINI can further refine their models with its more accurate contact predictions. In a set of 636 easy targets, DESTINI generates native-like structural models for 89% of targets versus 80% by TASSER. If one uses a higher TM-score > 0.5 as the cutoff, then DESTINI folded 478 (76%) targets versus 416 (66%) by TASSER. Overall, it is clear that the incorporation of a deep-learning algorithm into protein structure prediction significantly elevates the accuracy of computationally derived structural models.

We also compared the performance of DESTINI to several representative contact prediction methods. The benchmark set is composed of 66 domains from 50 targets evaluated during CASP12[9]. For this test, we removed from our training set all entries released after May 1st, 2016, the starting date of CASP12, and re-trained the network models with the reduced training set and a sequence library dated Feb 2016 for deriving the input features. For each target, we made the prediction for the full sequence with no domain partitioning performed. Domain partitioning was only performed for evaluation using the boundary provided by the assessors. Overall, DESTINI significantly outperforms the other methods. For the top *L*/2 medium or long range contacts, the mean precision of DESTINI is 70.1% versus 62.3% of RaptorX[10], the top ranked method in CASP12, 61.3% of DeepContact[11], which also employs a deep-learning algorithm, 60.7% of MetaPSICOV, the contact prediction leader in CASP11, and 42.8% of Gremlin, a standalone co-evolutionary analysis method. For the top *L*/5 medium or long range predictions, the mean precision is 78.8% for DESTINI, compared to 69.6%, 68.1%, 69.3%, and 47.1% for RaptorX, DeepContact, MetaPSICOV, and Gremlin.

Availability

Benchmark data sets and a DESTINI webserver are available at http://pwp.gatech.edu/cssb/destini.

1. Seemayer,S., Gruber, M. and Soding, J., *CCMpred--fast and precise prediction of protein residue-residue contacts from correlated mutations.* Bioinformatics, (2014). **30**(21): p. 3128-30.
2. Zhou,H. and Skolnick, J., *GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction.* Biophys J., (2011). **101**(8): p. 2043-52.
3. Gloor,G.B., et al., *Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions.* Biochemistry, (2005). **44**(19): p. 7156-65.
4. Altschul,S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Research, (1997). **25**(17): p. 3389-3402.
5. Jones,D.T., et al., *MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins.* Bioinformatics, (2015). **31**(7): p. 999-1006.
6. Zhou,H. and Skolnick, J., *Template-based protein structure modeling using TASSERVMT.* Proteins-Structure Function and Bioinformatics, (2012). **80**(2): p. 352-361.
7. Zhou,H. and Skolnick, J., *Ab initio protein structure prediction using chunk-TASSER.* Biophys. J., (2007). **93**: p. 1510--1518.
8. Skolnick,J. and Zhou, H., *Why Is There a Glass Ceiling for Threading Based Protein Structure Prediction Methods?* J Phys Chem B, (2017). **121**(15): p. 3546-3554.
9. Schaarschmidt,J., et al., *Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age.* Proteins, (2018). **86 Suppl 1**: p. 51-66.
10. Wang,S., et al., *Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model.* PLOS Comput Biol., (2017). **13**(1): p. e1005324.
11. Liu,Y., et al., *Enhancing Evolutionary Couplings with Deep Convolutional Neural Networks.* Cell Syst, (2018). **6**(1): p. 65-74.e3.

Distill for CASP13

M.Torrisi, M.Kaleel and G.Pollastri

School of Computer Science, University College Dublin, Ireland

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 residue contact maps submitted by Distill are predicted fully ab initio by an ensemble of 2D-Recursive Neural Networks trained on evolutionary features including correlated mutations.

Methods

Distill runs PSI-BLAST and hhblits against recent redundancy reduced versions of UniProtKB to generate multiple sequence alignments (MSA). The PSSM from the PSI-BLAST search is reloaded to search the PDB with PSI-BLAST for an initial guess at templates. MSA and templates are fed to our 1D prediction systems (all based on stacks of Bidirectional Recurrent Neural Networks and Convolutional Neural Networks): Porter[1,4,6,7] (secondary structure), PaleAle[4,6] (solvent accessibility), BrownAle[4] (contact density), Porter+[2] (structural motifs). All predictors use template information as an input alongside the sequence and MSA. The ab initio components of all predictors have recently been trained anew on sets of roughly 15,000 protein structures extracted from the PDB and should be considerably improved compared with the versions adopted at previous CASP editions.

1D predictions are combined into a structural fingerprint[4] (SAMD) which, alongside the PSSM, is used to find remote homologues in the PDB through 6 Smith-Waterman searches (PSSM and SAMD profile against PDB sequences and SAMD, with 3 different substitution matrices, plus 3 more searches against PDB PSSMs rather than sequences).

In parallel, residue contact maps with a contact threshold of 8Å are predicted by a newly trained system based on 2D-Recursive Neural Networks[5], and submitted to the RR category. Inputs for map prediction are: profiles from MSA; outputs from freecontact, CCMpred; selected 1D and 2D statistics from the MSA used. That is, the maps are always purely ab initio unlike Distill versions for previous CASP editions.

The 3D reconstruction, which is only conducted on C α 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 by substituting snippets of 3 to 9 amino acids extracted from the SAMD_list to quickly find a minimum of a potential function which rewards agreement with a set of desired constraints for the protein (see below); from the previous endpoint a low temperature refinement is run by substituting 9-mers from the conformation with 9-mers from the SAMD_list, and using the same potential function as above. The set of desired constraints driving the protein reconstruction is a weighted average of the distance maps of templates, interpolated, where templates are missing, with predicted ab initio maps as submitted to the RR category. That is, if no templates are found the reconstruction is purely based on our predicted contact map.

We run 30 reconstructions for each protein, which we rank by their weighed TM-scores against the template list and agreement with the predicted contact map. For the 5 top-ranked models we reconstruct the backbone with SABBAC, and the full atoms with Scwrl4. These are the models submitted to CASP.

Availability

The newest version of Distill is available at http://distilldeep.ucd.ie/casp/

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.
6. Mirabello, C. & Pollastri, G. (2013) Porter, PaleAle 4.0: high-accuracy prediction of protein secondary structure and relative solvent accessibility, *Bioinformatics*, **29**(16):2056-2058.
7. Torrisi, M, Kaleel, M & Pollastri, G. (2018) Porter 5: fast, state-of-the-art ab initio prediction of protein secondary structure in 3 and 8 classes, *bioRxiv* 289033; doi: https://doi.org/10.1101/289033

Contact prediction and de novo protein structure prediction using deep neural networks

H. Fukuda[1], T. Nakamura[1], Y. Yamamori[2], Y. Tsuchiya[2] and K. Tomii[1,2]

1 -Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, 2 - Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST)

k-tomii@aist.go.jp

For *de novo* protein structure prediction, we have developed a novel approach, to predict contacts in proteins, which combined unsupervised learning and supervised learning methods for predicting protein contacts with multiple sequence alignment (MSA) using a deep neural network (DNN)[1]. We used MSA as an input feature and predict the contacts through an extremely deep (over 60 layers) Convolutional Neural Network (CNN). Simultaneously, we assigned a weight to each sequence in MSA to eliminate "noisy" sequences automatically in a supervised manner. 3D-models for targets are obtained based on both contacts predicted by our method and secondary structures predicted by RaptorX-Property[2] with CONFOLD[3]. We submitted both the results of contact prediction and the 3D-models for targets.

Methods

Our method for *de novo* protein structure prediction consists of four steps: 1) Construct an input MSA for our DNN model by using -oa3m option of HHblits[*] with three iterations, 2) Predict secondary structures obtained by using RaptorX-Property and solvent accessibility for each residue using SCRATCH-1D[4], 3) Calculate a probability for being in contact of each residue pair using our DNN model based on MSA, the results of secondary structure and solvent accessibility, 4) Construct 3D-models with CONFOLD by using the top 2*L* predicted contacts (here, *L* corresponds to the sequence length for each target) and the predicted secondary structure. We selected and submitted the top 5 models constructed by CONFOLD.

Our DNN model was trained by 14680 proteins derived from the PISCES[5] cull pdb server. During the CASP experiment, we have continued to tune our models and their hyper parameters and used ensemble models for the late part of the competition.

1. Fukuda H, Tomii K. Deep Neural Network for Protein Contact Prediction by Weighting Sequences in a Multiple Sequence Alignment. bioRxiv doi: https://doi.org/10.1101/331926 2018
2. Wang S, Li W, Liu S, Xu J. Protein RaptorX-Property: a web server for protein structure property prediction. Nucleic Acids Res. 2016 Jul 8;44(W1):W430-5
3. Adhikari B, Bhattacharya D, Cao R, Cheng J. CONFOLD: Residue-residue contact-guided ab initio protein folding. Proteins. 2015 Aug;83(8):1436-49
4. Magnan C.N. and Baldi P. (2014). SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. Bioinformatics, vol 30 (18), 2592-2597
5. Wang G. and Dunbrack, R.L. Jr. PISCES: a protein sequence culling server. Bioinformatics, 19:1589-1591, 2003.

Improved covariation-based contact predictions enabled by deep learning and protein-specific data augmentations

S.M. Kandathil [1,2], J.G. Greener [1,2] and D.T. Jones[1,2*]

1 - University College London, Dept. of Computer Science, Gower Street, London WC1E 6BT, 2- The Francis Crick Institute, 1 Midland Road, London NW1 1AT.

*d.t.jones@ucl.ac.uk

The DMP (DeepMetaPSICOV) server implements a new deep learning-based contact prediction method that combines our previous MetaPSICOV[1] and DeepCov[2] methods. In addition it employs a number of novel data augmentation strategies to maximize the value of the limited 3D structure training data that is available.

For a target sequence of length L, input features are the 441 x L x L covariance inputs derived direct from the sequence alignments, as used in DeepCov; plus 58 channels of other covariation-based features as used in MetaPSICOV2 and a further two "housekeeping" channels to indicate sequence separation and to demarcate the sequence bounds. The 58 features include 1D features, such as sequence profiles and secondary structure probabilities from PSIPRED, that have been striped both vertically and horizontally to make square 2D matrices of rank L. This gives a total of 501 input feature channels.

The prediction model is a deep (77-layer), fully convolutional residual network model implemented with PyTorch. First, a Maxout layer[3] is used to reduce the dimensionality of the inputs (once again similar to DeepCov) to 64 feature channels, with 20% dropout then applied. Following this, each residual block consists of two 5x5 64-feature convolutional layers with instance normalisation applied throughout along with standard ReLU nonlinearities. A mixture of regular and dilated 5x5 filters are used, with the dilation rates increasing by a factor of 2 to a maximum of 64. Dilations are applied as a means to rapidly grow the receptive field of the network to encompass the whole protein input. Every convolutional layer (dilated or otherwise) uses padding on its inputs such that its input and output tensors have the same spatial dimensions (square matrices of rank L). The final layer is a 1x1 filter convolutional layer with sigmoid nonlinearity applied, where the final outputs represent the probability of each residue pair being in contact. Network parameters were trained using Adam optimisation[4] and a binary cross-entropy loss function. The final prediction for each target protein is the result of averaging predictions from five separately trained models with differing random seed values.

One key innovation in our approach is our extensive set of data augmentation approaches used for training. We augment our standard training set of 6729 protein alignments (200 of which are used as a validation set) using a variety of random transformations of the input feature maps e.g. sequence reversal, loop sampling and random input feature interpolation between shallow and deep alignments.

During inference, we generate alignments using a similar approach to that taken in MetaPSICOV2, with a few additions: first, we carry out basic homologous domain parsing based on an initial HHblits[5] search against the PDB70 database provided by the Soeding group. Then we run HHblits on the standard UNICLUST30 sequence data bank, and if an insufficient number of related sequences is found, we then run jackHMMER[6] on a custom non-redundant sequence data bank. This data bank is formed by merging the latest Uniref100 data bank together with the EBI metagenomics peptide data bank, at a redundancy threshold of 100% sequence identity. The sequence hits obtained by jackHMMER are then extracted and clustered using kClust[7] and aligned using MAFFT[8] to build MSAs in order to make a custom HHblits database for each target, against which a final run of HHblits is used to search. This strategy has proven beneficial in obtaining additional sequences in benchmarks, especially on targets with few sequence relatives, where often this provides as much as a twofold increase in alignment depth.

1. Buchan,D.W.A. & Jones,D.T. (2018) Contact predictions with the MetaPSICOV2 server in CASP12. *Proteins* **86**(Suppl 1), 78-83.
2. Jones,D.T. & Kandathil,S.M. (2018) High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. *Bioinformatics.* In press.
3. Goodfellow,I.J., Warde-Farley,D., Mirza,M., Courville,A. & Bengio,Y. (2013) Maxout networks. In: Sanjoy,D. and David,M., eds., Proceedings of the 30th International Conference on Machine Learning. *Proceedings of Machine Learning Research: PMLR*, 1319-1327.
4. Kingma,D.P. & Ba,J. (2014) Adam: A Method for Stochastic Optimization. *arXiv*:1412.6980.
5. Remmert,M., Biegert,A., Hauser,A. & Söding,J. (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9**, 173–175.
6. Eddy,S.R. (2011) Accelerated profile HMM searches. *PLoS Comp Biol* **7**(10), e1002195
7. Hauser,M., Mayer,C.E. & Söding,J. (2013) kClust: fast and sensitive clustering of large protein sequence databases. *BMC Bioinformatics* **14**, 248.
8. Katoh,K. & Standley,D.M. (2007) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30**(4), 772–780.

Using model quality assessment and template based and free docking in CASP13

Claudio Bassot[1,2], Petras Kundrotas[3] and Arne Elofsson[1,2]

1-Science for Life Laboratory, Stockholm University, Solna, Sweden, 2- Department of Biochemistry and Biophysics, Stockholm University, Stockholm, Sweden, 3- Center for Computational Biology snf Department of Molecular Biosciences, The University of Kansas, Lawrence, KS, United States

arne@bioinfo.se

The use of different quality assessment methods increases the probability of identifying the best protein model among all server models. Per each target, we selected the best model according to the quality assessment (QA) methods: Pcomb[1], ProQ4[2], ProQ3D[3], Pcons[4], and agreement with contacts predicted by PconsC4[5]. It was noted that for the majority of the targets in CASP13 all the QA methods selected similar models. For monomeric targets, the top ranked model for each of these methods was submitted as model one to five. However, in case of oligomeric targets, we applied a more extensive manual procedure described below.

Method

For each oligomeric target, we ran HHsearch [6] to identify oligomeric templates. In parallel, we selected the top scoring models with different QA predictors, and among them, we manually chose the models covering all potential representative folds. The models of the monomers were structurally aligned on the template. If the alignment of all the monomers results in a putative oligomeric complex, the structure was relaxed using the Rosetta package [7]. If the output of the minimization maintained the oligomeric fold, the oligomeric model was submitted.

When no templates were available or the structural alignment gave an unsatisfactory outcome, we performed a "contact prediction based" docking. Starting from a multiple sequence alignment (MSA) obtained by Jackhmmer [8] we predict the protein contacts using PconsC4 [5]. For the heteromeric complexes, the MSA of the monomers of each species were merged. The top scored solvent accessible contacts between the monomers, or in the case of homomer between residues further than 15Å in the model were used as restraints using the Haddock [9] docking software.

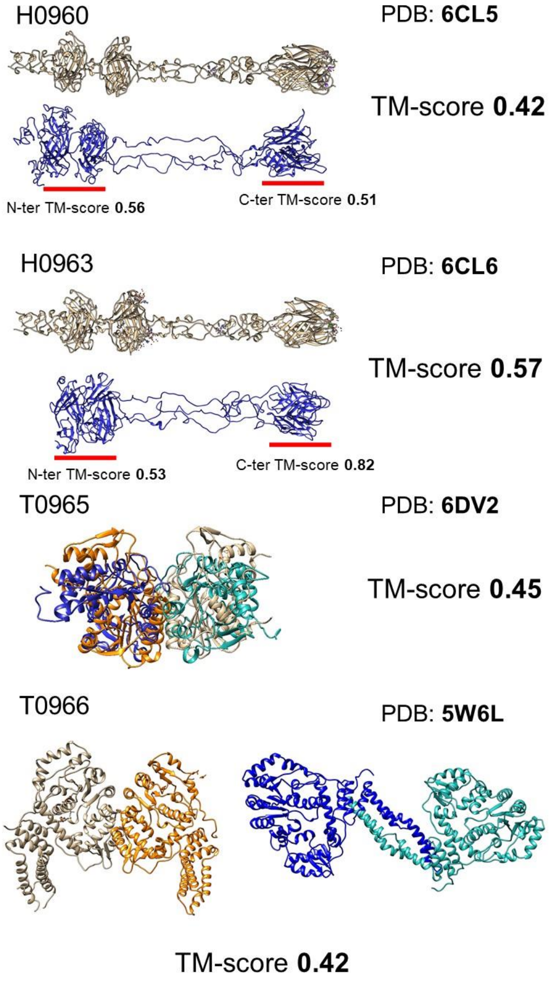
When no templates or good contacts prediction was available we performed a template-free docking using GRAMM[10] using the top scoring models.

Results

At the present date, five oligomeric targets are public on PDB, among these H0953 heteromer modelling failed because none of the servers modeled the s1 subunit with sufficient accuracy. For the remaining four targets we calculate the TM score between our best complex models and the PDB structures using MM-align[11] (Figure 1).

H0960 and H0963 were very similar targets: two elongated fibrillar proteins showing two globular domains with available templates. In both the cases, we docked the two globular domains based on the templates, and we connect them with loops generated with Profix package [12]. The TM score between our modes and the structures is respectively 0.42 for H0960 and 0.57 for H0963. Looking more in detail the structures, the two globular domains appear modelled significantly better than the connecting loops (Figure 1). The higher TM score for H0963 is mostly due to the fact that the N- terminal loops were not modelled and the better placement of the domains. A partially successful application of the template base docking is the homodimer T0965. Here, the dimerization interface was correctly predicted, but the structure presents an unexpected twist compared to our model. Finally, for T0966 a partial template and the contacts map suggested a wrong dimerization interface that may be an alternative to the one resolved in the PDB structure.

Figure 1. Four oligomeric targets with their respective PDB structures and their mutual TM score. In brown and orange the PDB structures, in blue and light blue our best model.



1. Larsson,P., Skwark,M.J., Wallner,B. & Elofsson,A. (2009). Assessment of global and local model quality in CASP8 using Pcons and ProQ. *Proteins Struct. Funct. Bioinforma.* **77**, 167–172.
2. Hurtado,D.M., Uziela,K. & Elofsson, A. (2018). Deep transfer learning in the assessment of the quality of protein models. *ArXiv180406281 Q-Bio*
3. Uziela,K., Menéndez Hurtado,D., Shu,N., Wallner,B. & Elofsson,A. (2017). ProQ3D: improved model quality assessments using deep learning. *Bioinforma. Oxf. Engl.*, **33**,1578–1580.
4. Wallner,B., Larsson,P. & Elofsson, A. (2007). Pcons.net: protein structure prediction meta server. *Nucleic Acids Res.* **35**, W369-374.
5. PconsC4: fast, free, easy, and accurate contact predictions. *bioRxiv*. Available at: https://www.biorxiv.org/content/early/2018/08/02/383133.
6. Söding,J. (2005). Protein homology detection by HMM–HMM comparison. *Bioinformatics*. **21**, 951–960.
7. 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., 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 Enzymol. **487**, 545–574.
8. Johnson,L.S., Eddy,S.R. & Portugaly,E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*.**11**, 431.
9. de Vries,S.J., van Dijk,M. & Bonvin,A.M.J.J. (2010). The HADDOCK web server for data-driven biomolecular docking. *Nat. Protoc*. **5**, 883–897.
10. Tovchigrechko,A. & Vakser,I.A. (2006). GRAMM-X public web server for protein–protein docking. *Nucleic Acids Res.* **34**, W310–W314.
11. Mukherjee,S. & Zhang,Y. (2009). MM-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming. *Nucleic Acids Res.* **37**, e83–e83.
12. JACKAL: A Protein Structure Modeling Package; Columbia University & Howard Hughes Medical, Institute: New York, (2002) http://honig.c2b2.columbia.edu/jackal/

FaeNNz - combining statistical potentials with consensus-based prediction of local model quality

G. Studer[1,2], C. Rempfer[1,2], J. Haas[1,2], R. Gumienny[1,2] and T. Schwede[1,2]

1 - Biozentrum, University of Basel, 2 - SIB, Swiss Institute of Bioinformatics

gabriel.studer@unibas.ch

Computational protein structure modeling methods, and in particular comparative modeling, have established themselves as valuable complement for structural analysis when experimental data is missing. While such methods have matured into stable and robust pipelines that can generate models for almost any protein automatically, the quality of the generated models can be highly variable and hard to predict in the absence of experimental observables. This is a major concern from an application perspective as the suitability of a model for a specific application directly depends on its quality, hence the importance of quality estimation methods. Currently, the most accurate QE methods rely on consensus information by assessing the variability in an ensemble of models, assuming that correct structural features will tend to be more conserved. However, this approach is only applicable if several alternative models are available. In contrast, knowledge based statistical methods can be applied to single models by comparing structural features in the model with those obtained from statistical distributions derived from high quality experimental structures. An example of the latter approach is QMEAN[1] developed in our group.

Methods

QMEAN uses statistical potentials of mean force and the consistency of a model with structural features predicted from sequence to generate quality estimates on a global and local scale.

Recently, QMEAN has been extended by the DisCo score. DisCo assesses the consistency of observed interatomic distances in the model with ensemble information extracted from experimentally determined protein structures that are homologous to the target sequence. In case many close homologous structures exist, DisCo is expected to be very accurate. However, if few or no close homologous templates can be identified, DisCo does not contain sufficient information for scoring models. In order to combine the ability of statistical potentials to score individual models with the power of DisCo in cases with sufficient template information, we use a random forest approach to optimally weigh the two components and derive a combined score for accurate local quality estimates: QMEANDisCo.

Since late 2017, QMEANDisCo is the default scoring method employed by the SWISS-MODEL server[2]. Within the CAMEO experiment[3], it exhibits excellent performance in detecting local errors in models of correct overall fold, as they are typically built in classical comparative modeling approaches. However, prediction performance significantly decreases when confronted with more diverse data as they are generated in CASP.

In this work, we revisited the QMEANDisCo approach and additionally introduced several low resolution terms, which are expected to assess the general overall fold. The score combination has been altered to use feedforward neural networks. Different compositions of training data originating from CASP12 and CAMEO have been evaluated to optimize local quality prediction performance on more diverse test sets. The resulting method has to be considered experimental and carries the working title FaeNNz (After a traditional recipe from the swiss alps: Faenz).

Results

Same as QMEANDisCo, FaeNNz combines the convenience of taking a single model as input with the predictive power of consensus approaches by combining statistical potentials with distance constraints from homologous templates. Using a 5-fold cross validation, FaeNNz exhibits comparable performance to QMEANDisCo on CAMEO data. This has been confirmed by subsequent blind evaluations by

CAMEO, where FaeNNz is registered as a test server. On CASP12 data, significant improvements have been observed. One interesting aspect of this ongoing research is the dependency of prediction performance on the underlying training data.

1. Benkert, P., Biasini, M., Schwede, T. (2011). Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics*. **27**(3), 343-50.
2. Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., de Beer, T.A.P., Rempfer, C., Bordoli, L., Lepore, R., Schwede, T. (2018) SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res*. **46**(W1), W296-W303.
3. Haas, J., Barbato, A., Behringer, D., Studer, G., Roth, S., Bertoni, M., Mostaguir, K., Gumienny, R., Schwede, T. (2018) Continuous Automated Model EvaluatiOn (CAMEO)complementing the critical assessment of structure prediction in CASP12. *Proteins.* **86**, 387-398.

Predicting protein inter-residue contacts using composite likelihood maximization and deep learning

Haicang Zhang[1,2], Qi Zhang[1,2], Fusong Ju[1,2], Jianwei Zhu[1,2], Shiwei Sun[1,2], Yujuan Gao[3], Ziwei Xie[1], Minghua Deng[3], Wei-Mou Zheng[*,4], and Dongbo Bu*[1,2]

1 - Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China; 2 - University of Chinese Academy of Sciences, Beijing, China; 3 - Center for Quantitative Biology, School of Mathematical Sciences, Center for Statistical Sciences, Peking University, Beijing, China; 4 - Institute of Theoretical Physics, Chinese Academy of Sciences, Beijing, China

FALCON@ict.ac.cn

Accurate prediction of inter-residue contacts of a protein is important to calculating its tertiary structure. Analysis of co-evolutionary events among residues has been proved effective to inferring inter-residue contacts. The Markov random field (MRF) technique, although being widely used for contact prediction, suffers from the following dilemma: the actual likelihood function of MRF is accurate but time-consuming to calculate; in contrast, approximations to the actual likelihood, say pseudo-likelihood, i.e., the product of conditional probability of individual residues, are efficient to calculate but inaccurate. Thus, how to achieve both accuracy and efficiency simultaneously remains a challenge. In this study, we present such an approach (called clmDCA) for contact prediction. Unlike plmDCA using pseudo-likelihood, our approach uses composite-likelihood, i.e., the product of conditional probability of all residue pairs. Composite likelihood has been theoretically proved as a better approximation to the actual likelihood function than pseudo-likelihood. Meanwhile, composite likelihood is still efficient to maximize, thus ensuring the efficiency of clmDCA.

Methods

For a query protein, clmDCA predicts its inter-residue contacts through the following three steps. First, we construct multiple sequence alignment (MSA) for homologous proteins of the query protein. According to the MSA, the correlations among residues are disentangled using the composite likelihood maximization technique, and are subsequently explored to infer contacts among residues. The generated inter-residue contacts are further refined using a deep residual network. These steps are described in more details as follows:

1. Modeling MSA using Markov random field

We use a vector of variables $X = (X_1, X_2, \cdot\cdot\cdot, X_L)$ to represent a protein sequence in MSA with X_i representing position i of MSA. According to the maximum entropy principle, the probability that X takes a specific value x^m can be represented using Markov random field model[1].

2. Direct coupling analysis using composite likelihood maximization

The maximization of the actual likelihood of MRF model is inefficient since the calculation of partition function Z under multiple parameter settings is needed. To circumvent this difficulty, pseudo-likelihood was used as an approximation to the actual likelihood[2]. To better approximate the actual likelihood, we use composite likelihood instead of pseudo-likelihood[3].

The advantages of pairwise composite likelihood technique are two-folds: (1) Compared with pseudo-likelihood, pairwise composite likelihood is a better approximation to the actual likelihood. (2) The gradients of pairwise can be calculated in polynomial time. Thus, the pairwise composite likelihood approach achieves both accuracy and efficiency simultaneously.

3. Refining inter-residue contacts using deep residual network

To refine the predicted contacts by clmDCA, we fed them into a deep residual network[4] for denoising. Deep residual network has its advantages in the ease of training process and the capacity of considerably deep architecture as each layer learns a residual function with reference to the layer input rather than unreferenced functions. We also considered the 1D information of the query protein, including sequence profile, predicted secondary structure, solvent accessibility.

Results

We tested clmDCA on PSICOV[5] data set (containing 150 proteins) and CASP-11 data set (containing 85 proteins). On the CASP-11 dataset, our clmDCA outperforms plmDCA and other purely-sequence-based approaches. Take top L/10 predictions with the sequence separation threshold 6AA as an example. clmDCA achieved prediction precision of 0.83, which is higher than plmDCA (0.81), mfDCA (0.73) and PSICOV (0.77). On the CASP-11 dataset, the prediction accuracy of all these approaches are relatively low than those on the PSICOV dataset. This might be attributed to the difference in MSA quality. However, clmDCA still outperformed other approaches.

When equipped with deep learning technique for refinement, both plmDCA and clmDCA achieved better prediction accuracy. For example, on the CASP-11 dataset, plmDCA and clmDCA alone achieved prediction accuracy of only 0.54 and 0.57, respectively (sequence separation > 6AA; top L/10 contacts). In contrast, by applying the deep learning technique for refinement, the prediction accuracies significantly increased to 0.77 and 0.86, respectively. More importantly, the improvement of clmDCA (from 0.57 to 0.86) is considerably higher than that of plmDCA (from 0.54 to 0.77), suggesting that clmDCA results are more suitable for refinement using deep learning technique.

Availability

http://protein.ict.ac.cn/clmDCA/

1. Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T., and Weigt, M. (2011a). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, **108**(49), E1293–E1301.
2. Ekeberg, M., Lovkvist, C., Lan, Y., Weigt, M., and Aurell, E. (2013). Improved contact prediction in proteins: using pseudolikelihoods to infer potts models. *Physical Review E*, **87**(1), 012707.
3. Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society*. Series B (Methodological), pages 192–236.
4. Wang, S., Sun, S., Li, Z., Zhang, R., and Xu, J. (2017a). Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS computational biology*, **13**(1), e1005324.
5. Jones, D. T., Buchan, D. W., Cozzetto, D., and Pontil, M. (2012). PSICOV: precise struc- tural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, **28**(2), 184–190.

Protein Model Refinement via Iterative Molecular Dynamics Simulations

Lim Heo and Michael Feig

Department of Biochemistry and Molecular Biology, Michigan State University, MI 48824, USA

mfeiglab@gmail.com; http://feiglab.org

Molecular dynamics-based refinement methods have shown promising results in the previous CASP experiments, and they have become part of the mainstream in protein model refinement. In CASP12, we examined a protocol that exploited microsecond-scale MD simulations with an improved force field[1]. Based on this analysis, it seemed that MD-based refinement method may have reached a plateau but without quite reaching the ultimate goal of producing structures of near-experimental accuracy. To tackle the issue, we have carried out detailed analyses to identify the major bottlenecks that hinder further progress. Many refinement methods use restraints biasing to the initial model not to deteriorate much from the experimental structures. However, those restraints increase energy barriers toward the other states, and they only allow limited conformational sampling near the initial state. Moreover, since many structure prediction servers have adopted MD-based refinement steps as part of their modeling pipelines, it became harder to further refine models using similar methods.

During CASP13, we tested two MD-based refinement protocols. The main protocol is based on iterative MD simulations with wider flat-bottom harmonic restraints. This protocol allowed much broader sampling while still keep the conformational space exploration in the vicinity of the– experimental structure. The second protocol was analogous to our previous protocol[1,2] but using reduced simulation time. By default, the final model from the main protocol was submitted as “model 1”, while the final model from the second, established protocol replaced it in certain cases where we did not have confidence that the new protocol was successful. As a final step, our method locPREFMD[3] was applied to all of the submitted models.

Methods

The main protocol consists of iterative rounds of MD simulations with flat-bottom harmonic restraints, scoring with the most recent Rosetta energy function[4], and structure averaging of selected conformations. Three iterations of simulations were performed for CASP13 predictions. Before simulations were started, locPREFMD was applied to the initial models to remedy potential stereochemical problems. The resulting model was then used as the initial structure for the first iteration of sampling. In the first iteration, five independent 100 ns-long MD simulations were carried out. The sampled conformations were clustered based on Calpha-RMSD, and further lumped into a Markov-state model by considering that their kinetics have a minimum relaxation time longer than 20 ns. For the second and third iterations, different initial conformations were selected from cluster-averaged structures generated in the previous iteration. A cluster was selected for the next iteration unless it was used as an initial structure in the previous iterations, and up to 5 and 10 clusters were selected according to their size. For the second and third iterations, ten and twenty independent 50 ns-long MD simulations were carried out. Once the iterative sampling was complete, the sampled conformations were scored after local minimization by using Rosetta with the “minimize” executable using default parameters. A quarter of structures with the lowest Rosetta scores were selected and averaged to give the final model.

MD simulations were conducted with a modified CHARMM force field with explicit solvents. The latest CHARMM force field, c36m[5], was modified with lower energy barriers in backbone dihedral angles via an alternate CMAP cross-correlation torsion term to accelerate conformational transitions. In addition, hydrogen atoms were assigned heavier masses (3 a.m.u.) by re-distributing mass from hydrogen-attached heavy atoms so that MD simulations could be run with a 4-fs integration time step. Flat-bottom harmonic restraints were applied to every Calpha atom with respect to the initial structure. In the second and third iterations, the bias was applied with respect to the respective starting structures rather than the initial model. The flat-bottom potentials did not restrain sampling up to 4 Å away from the reference but became active when conformations deviated further.

The secondary protocol is analogous to our previous protocol used in the previous CASP experiments. It also begins with the initial model after applying locPREFMD. In this protocol, five independent 50 ns-long MD simulations were run with weak harmonic restraints on Calpha atoms. MD simulations were conducted with the standard c36m CHARMM force field with explicit solvents. To filter out poor conformations, sampled structures were locally minimized and scored by using Rosetta as in the main protocol. A quarter of structures was excluded, and the remaining structures were averaged to give the final model for this protocol.

The "model 1" submissions were taken either from protocol 1 or 2. By default, the final model from the main protocol was submitted as "model 1" except in the following cases where preliminary tests suggested that the main protocol would not perform as well: (1) larger targets having radii of gyration greater than 17 Å, (2) targets with highly unstable initial models, and (3) targets with putative ligands bound. We considered an initial model highly unstable initial model when sampled structures deviated from the initial model more than 1 Å for more than 99% of the simulation time. As a significant number of targets fell into those categories the "model 1" submissions came from the main protocol for only 19 out of the 29 total refinement targets.

MD simulations were carried out by using OpenMM[6]. Simulations were prepared and analyzed by using CHARMM[7] and the MMTSB toolset[8]. We used in-house Python script based on MSMbuilder[9] and the MDTraj[10] library for the clustering. Rosetta (version 3.9) was used to evaluate the sampled conformations.

1. Heo,L., Feig,M. (2018) What makes it difficult to refine protein models further via molecular dynamics simulations? *Proteins* **86**(Suppl 1), 177-188.
2. Heo,L., Feig,M. (2018) PREFMD: a web server for protein structure refinement via molecular dynamics simulations. *Bioinformatics* **34**(6), 1063–1065.
3. Feig,M. (2016) Local Protein Structure Refinement via Molecular Dynamics Simulations with locPREFMD. *J. Chem. Inf. Model.* **56**(7), 1304–1312.
4. Park,H., et al. (2016) Simultaneous optimization of biomolecular energy function on features from small molecules and macromolecules. *J. Chem. Theory Comput.* **12**(12), 6201–6212.
5. Huang.J., et al. (2016) CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nature Methods* **14**(1), 71–73.
6. Eastman,P., et al. (2017) OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* **13**(7), e1005659–17.
7. Brooks,B.R., et al. (2009) CHARMM: the biomolecular simulation program. *J. Comput. Chem.* **30**(10), 1545–1614.
8. Feig,M., Karanicolas,J., Brooks,C.L. III (2004) MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology. *J. Mol. Graph. Model.* **22**(5), 377–395.
9. Harrigan,M.P., et al. (2017) MSMBuilder: Statistical Models for Biomolecular Dynamics. *Biophys. J.* **112**(1), 10–15.
10. McGibbon,R.T., et al. (2015) MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **109**(8), 1528–1532.

Combining pyDock ab initio docking and template-based modeling for the CASP13-CAPRI Challenge

Mireia Rosell, Luis Angel Rodríguez-Lumbreras, Miguel Romero, Lucía Díaz,
and Juan Fernández-Recio

Barcelona Supercomputing Center (BSC), IBMB-CSIC

juanf@bsc.es

Structural modeling of oligomeric proteins can largely benefit from *ab initio* docking procedures, as part of an integrative approach including template-based data, experimental information on the interface residues, and symmetry restraints. To evaluate this, we have participated, both as predictors and as scorers, in all the 22 targets proposed for the joint challenge between CASP13 and CAPRI Round 46, comprising ten homo-dimers (A:A), four hetero-dimers (A:B), one homo-tetramer (A4), one hetero-tetramer (A2:B2), one homo-octamer (A8), one (likely) trimer (A3), one hetero-18-mer (A6:B6:C6), and one homo-dimer of 5-domain monomers (A_5D:A_5D). The latter complex was also used to define two more targets with additional SAXS data and cross-linking information, respectively.

Methods

Generation of docking and template-based models

In general, we used *ab initio* docking to model the oligomeric targets. We used as starting subunits the best prediction from ZHANG, ROSETTA, and QUARK CASP-hosted servers (according to their order of submission). Then, for each pair of modelled subunits (usually three pairs per target), FTDock[1] (with electrostatics and 0.7 Å grid resolution) and ZDOCK 2.1[2] were used to generate 10,000 and 2,000 rigid-body docking poses, respectively, and the resulting models were merged in a single pool. In homo-oligomers, docking poses not satisfying the expected symmetry (e.g. C2 for homo-dimers, C3 for homo-timers, etc.) were removed.

In cases with available homologous oligomeric templates, models were also built based on such templates. First, we used the top five released predictions from the ZHANG, ROSETTA, QUARK, MULTICOM-CONSTRUCT and RAPTOR Deep Modeller CASP-hosted servers as starting models of the individual monomers of each target. Then, we used BLAST to search for suitable oligomeric templates for each complex, and also extracted oligomeric templates from the above mentioned servers participating in CASP. Templates were clustered to remove redundant structures. The monomeric models were superimposed onto the corresponding subunits of each selected template.

In some targets, the integration of template-based and ab initio docking was essential to produce feasible models. This is the case of target T146 (A2:B2) in which the homo-dimer interfaces were modelled based on available templates, and the heteromeric interfaces by docking. Similarly, in the hetero-18-mer target T159 (A6:B6:C6), the homo-hexameric rings were modelled based on available templates, and then the interaction between the rings were modelled by FTDock docking and/or based on templates.

Scoring of oligomeric models

We scored the above described oligomeric models with pyDock,[3] sorting them according to the total binding energy of all possible interfaces. In target T146, available Cryo-EM information[4] was used to filter the docking results. The number of available templates and their reliability determined the percentage of template-based complex models included in the final 5 submitted models (10 for CAPRI). Finally, we eliminated the redundant predictions and minimized the final ten selected docking models.

In the scorers experiment, we eliminated all the docking models with a percentage of secondary structure significantly lower than the one observed in the corresponding set of structures previously selected as predictors. Models with more than 250 clashes (i.e. intermolecular pairs of atoms closer than 3 Å) were also removed. Then, the same protocol used in predictors was applied to score the docking models (favoring models structurally similar to reliable available templates).

Dimerization of a multi-domain protein: a challenging case

For target T149, involving the dimerization of a 5-domain protein, a specific modeling strategy was devised, since the challenge was not only to model the dimer orientation but also to describe the assembly of the 5 different domains within each monomer. The domains were modelled based on the best monomeric prediction submitted to CASP by QUARK server. The intermolecular orientation between the first domain from each monomer was modelled by superimposing them on an available template. Then the remaining individual domains were docked to each other by using FTDock, describing all possible combinations (domain 1 vs. domain 2, domain 2 vs. domain 2, etc.). The best-scoring docking pairs were sequentially selected to grow the oligomeric model in a hierarchical manner, avoiding clashes and imposing intramolecular domain-domain restraints derived from the inter-domain linkers with pyDockTET.[5] When SAXS data was made available (constituting target T150), models were re-scored with pyDockSAXS.[6] Finally, when cross-linking data was released (target T151), the number of residue pairs satisfying these experimental contacts was also evaluated.

Availability

The pyDock[5] program is available for academic use as a GNU/Linux binary and as a web server (https://life.bsc.es/pid/pydock/).

1. Gabb, H.A., Jackson, R.M. & Sternberg, M.J. Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J. Mol. Biol.* **272,** 106–120 (1997).
2. Chen, R., Li, L. & Weng, Z. ZDOCK: an initial-stage protein-docking algorithm. *Proteins* **52,** 80–87 (2003).
3. Cheng, T.M.-K., Blundell, T.L. & Fernandez-Recio, J. pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. *Proteins* **68,** 503–515 (2007).
4. Ekiert, D.C., Bhabha, G., Isom, G.L., Greenan, G., Ovchinnikov, S., Henderson, I.R., Cox, J.S., Vale, R.D. Architectures of Lipid Transport Systems for the Bacterial Outer Membrane. *Cell* **169**, 273-285 (2017)
5. Cheng, T. M.-K., Blundell, T. L. & Fernandez-Recio, J. Structural assembly of two-domain proteins by rigid-body docking. *BMC Bioinformatics* **9,** 441 (2008).
6. Jiménez-García, B., Pons, C., Svergun, D.I., Bernadó, P. & Fernandez-Recio, J. pyDockSAXS: protein-protein complex structure by SAXS and computational docking. *Nucleic Acids Res.* **43**, W356-361 (2015).

GAPF_LNCC: an automated method for protein structure prediction with a multiple minima genetic algorithm

F.L. Custódio and L.E. Dardenne

Laboratório Nacional de Computação Científica, Petrópolis-–RJ, Brasil

flc@lncc.br

We built a fully automated de novo PSP method, that can be easily integrated into a web server, (GAPF_LNCC workflow). At the time of CASP13 we did not have the hardware necessary to host a server, therefore we participated in the all-groups category. The workflow is based on two ideas: (i) the use of experimental information available, in the form of residue-residue contact prediction, secondary structure prediction, and fragments, and (ii) a multiple minima genetic algorithm for conformational search. We employ a coarse-grained representation where all backbone atoms are explicit, with the side chains modeled as a single superatom. The scoring function combines some physically realistic potential with knowledge-based terms to promote hydrogen bonding and secondary structure organization. Global optimization is carried out by the multiple-minima genetic algorithm (GA) and no further refinement is performed. Selection of the models is then done by means of structural redundancy filtering and energy pruning. The GAPF_LNCC workflow was applied to 56 targets. Targets with less than 200 residues were prioritized. A template-based de novo strategy was used starting from target T0986s1 when suitable templates were found.

Methods

Most accessory programs and tools are run locally on a desktop PC, RaptorX contacts prediction being the only exception. The conformational search step runs at the Santos Dumont cluster. Our workflow starts with (1) secondary structure prediction by PSIPRED[1] followed by (2) domain prediction by INTERPROSCAN[2]. (3) Residue-residue (RR) contacts prediction is made by the RaptorX server[3] and when a prediction cannot be obtained in time DeepCov[4] is executed locally. (4) Fragment libraries are created with Profrager[5] (https://www.lncc.br/sinapad/Profrager/), and fragments are selected using the secondary structure prediction, residue-residue contacts score, in addition to the local sequence similarities from a culled database of 34,750 chains from experimental structures.

The (5) conformational search carried out by GAPF[6] employs a genetic algorithm (GA) with seven genetic operators including Ramachandran based mutations[7] and fragment insertion. The GA methodology uses a scoring function with a proper dihedral, steric repulsion, hydrophobic compaction, hydrogen bonding formation[8], cooperative hydrogen bonding[9] and RR contacts[10] terms. GAPF employs a phenotype- based crowding mechanism for the maintenance of useful diversity within the populations, which has been shown to result in increased performance and to grant the algorithm multiple solution capabilities. For each target, at most 100 independent runs of the GA ware performed (dependant on time restraints), each population contains 200 individuals, resulting in 20,000 structures. These results undergo a (6) structural redundancy filter and the overall top five structures, ranked by energy, proceeded to the next steps. If the target is split into domains the (7) final structures are assembled with another run of GAPF where the initial population is seeded with 50 random combinations of structures from step 6. The resulting structures undergo another round of filtering identical to step 6. (8) Side chains of the select structures are reconstructed using SCWRL4[11]. And finally, the files are (9) formatted according to CASP guidelines, including (10) filling the temperature column of the PDB files with the confidence in the prediction (0-1, where 0 is the worst). Starting from target T0986s1 templates were sought using HHblits[12] those found with probabilities larger than 70% are used to seed the initial populations of the genetic algorithm.

Availability

Fragment library generation with Profrager is available at https://www.lncc.br/sinapad/Profrager/. Other tools are freely available from their authors. Some steps in the protocol were carried out with experimental versions of our software that will be made available by contacting the authors. The fully functional web server should open by mid-2019. Acknowledgment: FAPERJ (grant E-26/010.001229/2015), Santos Dumont supercomputer (http://www.sdumont.lncc.br/). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

1. McGuffin, L. J., Bryson, K., & Jones, D. T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics*, *16*(4), 404-405.
2. Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., & Lopez, R. (2005). InterProScan: protein domains identifier. *Nucleic acids research*, *33*(suppl 2), W116-W120.
3. Wang, S., Sun, S., & Xu, J. (2018). Analysis of deep learning methods for blind protein contact prediction in CASP12. *Proteins: Structure, Function, and Bioinformatics*, *86*, 67–77. http://doi.org/10.1002/prot.25377
4. Jones, D. T., & Kandathil, S. M. (2018). High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. *Bioinformatics*. http://doi.org/10.1093/bioinformatics/bty341
5. Santos, K. B., Trevizani, R., Custodio, F. L., & Dardenne, L. E. (2015, January). Profrager Web Server: Fragment Libraries Generation for Protein Structure Prediction. In*Proceedings of the International Conference on Bioinformatics & Computational Biology (BIOCOMP)* (p. 38). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).
6. Custódio, F. L., Barbosa, H. J., & Dardenne, L. E. (2014). A multiple minima genetic algorithm for protein structure prediction.*Applied Soft Computing*, *15*, 88-99.
7. Santos, K. B., Custódio, F. L., Barbosa, H. J., & Dardenne, L. E. (2015, August). Genetic operators based on backbone constraint angles for protein structure prediction. In*Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2015 IEEE Conference on* (pp. 1-8). IEEE.
8. Rocha, G. K., Custódio, F. L., Barbosa, H. J. C., & Dardenne, L. E. (2015, August). A multiobjective approach for protein structure prediction using a steady-state genetic algorithm with phenotypic crowding. In *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2015 IEEE Conference on* (pp. 1-8). IEEE.
9. Levy-Moonshine, A., Amir, E. A. D., & Keasar, C. (2009). Enhancement of beta-sheet assembly by cooperative hydrogen bonds potential. *Bioinformatics*, *25*(20), 2639-2645.
10. Santos, K. B., Rocha, G. K., Custodio, F. L., Barbosa, H. J. C., & Dardenne, L. E. (2017). Improving de novo protein structure prediction using contact maps information. In 2017 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB) (pp. 1–6). IEEE. http://doi.org/10.1109/CIBCB.2017.8058535
11. Krivov, G. G., Shapovalov, M. V., & Dunbrack, R. L. (2009). Improved prediction of protein side-chain conformations with SCWRL4. *Proteins: Structure, Function, and Bioinformatics*, *77*(4), 778-795.
12. Remmert, M., Biegert, A., Hauser, A., & Söding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature Methods, 9(2), 173–175. http://doi.org/10.1038/nmeth.1818

Predicting protein structure in CASP13 challenge

S. Grudinin[1], M. Karasikov[2], and G. Pagès[1]

1 - Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France, 2 - Department of Computer Science, ETH Zurich, Zurich, 8092, Switzerland

Sergei.Grudinin@inria.fr

Protein structure prediction an important and still unsolved problem in structural bioinformatics. It involves predicting structure of individual domains, or their complexes. In some cases additional information about the targets is available, so integrative modeling approaches can be used. We have applied methods and algorithms developed in our team for structure prediction of several categories of targets in the CASP13 protein modeling challenge. These included regular, assembly, SANS-, SAXS-, and X-link- assisted targets.

Methods

Structures of the monomers were selected from the server submissions based on their SBROD scores[1]. Structures of the multimers were produced with exhaustive sampling using Hex[2] and Sam[3] rigid-body fast Fourier transform-accelerated docking engines starting from the 50 best stage-2 server predictions. For heterooligomeric assemblies we used Hex, for homooligomeric assemblies we used Sam, and for the mixed stoichiometries we used a combination of two. The docking results were rescored with the SBROD scoring function. This function was trained using slightly different training sets for the Grudinin and SBROD groups. More precisely, the SBROD group was using potential trained on rounds 5-12 CASP server submissions enhanced with normal mode analysis-based decoys[6,7], and the Grudinin group was additionally using in the training set human submissions from the same CASP rounds.

Structures of the SAXS- and SANS- assisted targets were selected using Pepsi-SAXS[4] and Pepsi-SANS[5] methods, respectively. For the multimeric targets, we were using scoring in the polynomial space. Chi2 ranking was used by the Grudinin group and ranking based on the SBROD scores was used in the SBROD group.

Finally, cross-linking (XL) targets were optimized using gradient descent with respect to the XL harmonic energy term. For the multimeric targets, we additionally scored the binding poses according to the XL energy contribution. In the SBROD group, the final model ranking was performed using the SBROD scores.

Results

We applied the described methods to the structure predictions (TS category) targets in the CASP13 challenge including regular, assembly, SANS-, SAXS-, and X-link- assisted targets.

Availability

Our methods will be made available on our website at https://team.inria.fr/nano-d/software/.

1. Karasikov M., Pages G., Grudinin, S (2018) Smooth orientation-dependent scoring function for coarse-grained protein quality assessment. In revision.
2. Ritchie D.W., Kemp, G.J.L. (2000) Protein Docking Using Spherical Polar Fourier Correlations. *Proteins: Struct. Funct. Genet.* **39**: 178-194.
3. Ritchie D.W., Grudinin, S (2016) Spherical polar Fourier assembly of protein complexes with arbitrary point group symmetry. *J. Appl. Cryst*. **12** :1019-1028.
4. Grudinin,S., Garkavenko,M., & Kazennov,A. (2017). Pepsi-SAXS : an adaptive method for rapid and accurate computation of small-angle X-ray scattering profiles. *Acta Cryst D*, **73**, 449 – 464.

5. https://team.inria.fr/nano-d/software/pepsi-sans/
6. Hoffmann,A., & Grudinin,S. (2017). NOLB: Nonlinear rigid block normal-mode analysis method. *J. Chem. Theory Comput.*, **13**, 2123-2134.
7. Neveu,E., Popov,P., Hoffmann,A., Migliosi,A., Besseron,X., Danoy,G., Bouvry,P. & Grudinin, S, (2018), RapidRMSD: Rapid determination of RMSDs corresponding to motions of flexible molecules. *Bioinformatics*, **34**, 2757–2765.

Combining FFT-accelerated docking with a coarse-grained orientation-dependent potential for template-free modeling of protein complexes in CASP13 / CAPRI round 46 challenges

S. Grudinin[1] and M. Karasikov[2]

1 - Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France, 2 - Department of Computer Science, ETH Zurich, Zurich, 8092, Switzerland

Sergei.Grudinin@inria.fr

Protein docking is an important and still unsolved problem in structural bioinformatics. Docking approaches typically involve two steps - sampling the conformational space and scoring the obtained solutions. Based on our previous CASP/CAPRI experience, we used the best 50 server submissions for the sampling step. The challenge here was to develop a robust scoring methodology able to identify near-native contacts between docking partners, if the corresponding structures are predicted by very different methods. To do so, we utilized our recent coarse-grained potential[1], which is insensitive to the positions of the protein side chains.

Methods

For the sampling step, we used Hex[2] and Sam[3] rigid-body fast Fourier transform-accelerated docking engines to generate a vast amount of putative binding poses. We used the 50 best stage-2 server predictions, as ranked by the SBROD model quality assessment function[1], as starting docking models, and performed 1275 cross-docking runs for heterooligomers and 50 runs for homooligomers. For heterooligomeric assemblies we used Hex, for homooligomeric assemblies we used Sam, and for the mixed stoichiometries we used a combination of two.

The docking results were rescored with the SBROD scoring function. It uses only the backbone protein conformation, and hence it can be applied to scoring coarse-grained protein models. SBROD deduces its scoring function from a training set of protein models (server submissions from previous CASP rounds). The SBROD scoring function is composed of four terms related to different structural features. These are relative residue-residue orientations, contacts between backbone atoms, hydrogen bonds and solvent-solvate interactions. In order to identify the best contacts between the protein subunits, the final score was the difference between the score of the complex and its subunits.

Results

The docking method was applied to all multimers in the CASP13/CAPRI 46 experiment. It is relatively fast, as was run on a single laptop for all the targets. The SBROD scoring function was not specifically optimized to predict protein complexes, however, our preliminary tests demonstrated its ability to do so.

Availability

The SBROD scoring function inplimented in C++ and Python is freely available at https://gitlab.inria.fr/grudinin/sbrod and supported on Linux, MacOS, and Windows.

1. Karasikov M., Pages G., Grudinin, S (2018) Smooth orientation-dependent scoring function for coarse-grained protein quality assessment. In revision.
2. Ritchie D.W., Kemp, G.J.L. (2000) Protein Docking Using Spherical Polar Fourier Correlations. Proteins: Struct. Funct. Genet. 39: 178-194.
3. Ritchie D.W., Grudinin, S (2016) Spherical polar Fourier assembly of protein complexes with arbitrary point group symmetry. *J. Appl. Cryst*. **12** :1019-1028.

Modeling of proteins and their complexes guided by small-angle scattering profiles

S. Grudinin[1], A. Martel[2], S. Prevost[2], A. Hoffmann[1], and G. Dhar[1]

1 - Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France, 2 - Institut Laue-Langevin, Grenoble, 38000, France

Sergei.Grudinin@inria.fr

While crystallography has been providing atomic-resolution structures of biomolecules for over half a century, the real challenge of today's biophysicists is to correlate molecules' structure and dynamics in solution with their function. Small-angle scattering (SAS) is the fundamental techniques for structural studies of biological systems in solution. Thanks to advances in instrumentation and data analysis software, small-angle X-ray scattering (SAXS), complemented by other methods, is becoming very popular in structural biology. Over the years, a number of computational tools have been developed for the analysis of SAXS curves, calculation of theoretical profiles and low-resolution reconstruction of model shapes. Many efforts have been spent to reduce the running time of these tools without degrading the quality of their approximations. The number of Bio-SAXS publications exploded as a result of this effort. Comparatively, the lack of user-friendly analysis tools has hindered the development of small-angle neutron scattering (SANS), more complex but providing more information.

Methods

Recently, we developed SAXS and SANS packages called Pepsi-SAXS[1,2], and Pepsi-SANS[3], correspondingly. Pepsi-SAXS is a very efficient method that calculates small angle X-ray scattering profiles from atomistic models. It is based on the multipole expansion scheme and is significantly faster with the same level of precision compared to CRYSOL, FoXS and other methods. The method was systematically validated using an excessive set of over fifty models collected from the BioIsis and SASBDB databases. We have later extended it for neutron scattering applications[3].

One of the challenges of structural biology is flexible fitting of atomistic models into small-angle scattering profiles. Very recently, we designed a computational scheme that uses all-atom nonlinear normal modes[4,5] as a low-dimensional representation of the protein motion subspace and optimizes protein structures guided by the SAXS and SANS profiles. For example, in the CASP12 exercise, this scheme obtained best models for 3 out of 9 SAXS-assisted targets[6]. Overall, this flexible fitting scheme typically allows a significant improvement of the goodness of fit to experimental profiles in a very reasonable computational time.

Another challenge in the field is data-assisted protein docking. We have designed a scheme for SAXS- and SANS- assisted rescoring of docking predictions. This was made possible thanks the polynomial representation of partial scattering amplitudes for each of the docking partners. The scheme is very computationally efficient, it allows explicit representation of the hydration shell, and computes around 100,000 of Chi2 values per minute on a standard laptop for a mid-size protein complex.

Results

We applied the presented scheme to all SAXS- and SANS- assisted targets from the CASP13 challenge. For single-subunit targets, we used iterative optimization of Chi2 goodness of fit to experimental values while moving models along 10 slowest nonlinear modes. For protein complexes, we generated ~200,000 docking poses using Hex[7] for heterooligomers and Sam[8] for homooligomers. As starting models, we used 50 best server predictions, as was computed with our SBROD scoring model[9]. For each docking pose, we then computed corresponding Chi2 values using the polynomial representation of scattering coefficients. Finally, we clustered the solutions, and did the final optimization of the obtained complexes along normal modes. The final models were ranked by Chi2 (Grudinin group) or by the SBROD score (SBROD group).

Availability

The methods are available at http://team.inria.fr/nano-d/software/. A SAXS/SANS server is available at http://pepsi.app.ill.fr.

1. Grudinin,S., Garkavenko,M., & Kazennov,A. (2017). Pepsi-SAXS : an adaptive method for rapid and accurate computation of small-angle X-ray scattering profiles. *Acta Cryst D*, **73,** 449 – 464.
2. https://team.inria.fr/nano-d/software/pepsi-saxs/
3. https://team.inria.fr/nano-d/software/pepsi-sans/
4. Hoffmann,A., & Grudinin,S. (2017). NOLB: Nonlinear rigid block normal-mode analysis method. *J. Chem. Theory Comput.*, **13**, 2123-2134.
5. https://team.inria.fr/nano-d/software/nolb-normal-modes/
6. Tamò,G.E., Abriata,L.A., Fonti,G., Dal Peraro,M., (2018). Assessment of data-assisted prediction by inclusion of crosslinking/mass-spectrometry and small angle X-ray scattering data in the 12th Critical Assessment of protein Structure Prediction experiment. *Proteins: Struct., Funct., Bioinf*. **86**, 215–227.
7. Ritchie,D.W., & Kemp,G.J.L. (2000). Protein docking using spherical polar Fourier correlations. *Proteins: Struct., Funct., Bioinf*. **39**, 178-194.
8. Ritchie,D.W., & Grudinin,S. (2016). Spherical polar Fourier assembly of protein complexes with arbitrary point group symmetry. *J. Appl. Cryst*. **49**, 158-167.
9. Karasikov,M., Pages,G., & Grudinin,S. Smooth orientation-dependent scoring function for coarse-grained protein quality assessment. In revision.

Recurrent geometric networks for contact prediction

M. AlQuraishi

Department of Systems Biology, Harvard Medical School, 200 Longwood Ave., Boston, MA, USA

alquraishi@hms.harvard.edu

Recurrent geometric networks[1] (RGNs) are used to predict contact maps of protein structures. Intra-protein residue-residue distances are computed from raw structures predicted by RGNs, and monotonically transformed into contact probabilities. No templates were used during prediction, and no manual intervention was carried out. The raw amino acid sequence and position-specific scoring matrix (PSSM) of each protein was used as input.

Methods

We use the newly described RGNs[1] to perform contact predictions. RGNs are end-to-end differentiable models of protein structure which combine four ideas: (1) the adoption of a recurrent neural network architecture to encode the internal representation of protein sequence, (2) the parameterization of (local) protein structure by torsional angles, which provides a way to reason over protein conformations without violating the covalent chemistry of protein chains, (3) the coupling of local protein structure to its global representation via recurrent geometric units, and (4) the use of a differentiable loss function to capture deviations between predicted and experimental structures. RGNs replace conventional protein structure prediction pipelines with neural network primitives, making predictions of 3D structures directly from sequences + PSSMs without use of explicit templates, energy functions, or conformational sampling. As a result, the average prediction time for a new protein structure is 10 milliseconds.

We use a previously reported RGN model trained on the ProteinNet12 dataset, which contains all protein sequences and structures publicly available from UniParc + JGI metagenomes and the PDB, respectively, prior to the start of CASP12 (May 1st, 2016). We apply this model to CASP13 sequences and extract contact maps from the RGN-predicted 3D structures.

Availability

Source code for training new RGN models, as well as the trained RGN model used in this experiment, will be available shortly on GitHub.

1. AlQuraishi,M. (2018). End-to-end differentiable learning of protein structure. *bioRxiv*.
2. AlQuraishi,M. (2018) https://github.com/aqlaboratory/proteinnet, GitHub.

Recurrent geometric networks combined with Rosetta FastRelax Protocol

M. AlQuraishi

Department of Systems Biology, Harvard Medical School, 200 Longwood Ave., Boston, MA, USA

alquraishi@hms.harvard.edu

Recurrent geometric networks[1] (RGNs) are first used to predict backbone-only protein structures (atomic coordinates). Side-chains are then added and backbones are refined using the Rosetta FastRelax protocol. No templates were used during prediction, and no manual intervention was carried out. The raw amino acid sequence and position-specific scoring matrix (PSSM) of each protein was used as input.

Methods

We use the newly described RGNs[1] to perform contact predictions. RGNs are end-to-end differentiable models of protein structure which combine four ideas: (1) the adoption of a recurrent neural network architecture to encode the internal representation of protein sequence, (2) the parameterization of (local) protein structure by torsional angles, which provides a way to reason over protein conformations without violating the covalent chemistry of protein chains, (3) the coupling of local protein structure to its global representation via recurrent geometric units, and (4) the use of a differentiable loss function to capture deviations between predicted and experimental structures. RGNs replace conventional protein structure prediction pipelines with neural network primitives, making predictions of 3D structures directly from sequences + PSSMs without use of explicit templates, energy functions, or conformational sampling. As a result, the average prediction time for a raw protein structure from RGNs is 10 milliseconds.

We use a previously reported RGN model trained on the ProteinNet12 dataset, which contains all protein sequences and structures publicly available from UniParc + JGI metagenomes and the PDB, respectively, prior to the start of CASP12 (May 1st, 2016). We apply this model to CASP13 sequences. Raw RGN predictions, which contain only backbone atoms, are then refined using the Rosetta FastRelax protocol, which adds side chain atoms and refines the overall structure. For each sequence, we carry out 1,000 separate runs using FastRelax, and pick the five structures whose TM score relative to the raw RGN prediction is highest. On average the total amount of FastRelax computation per structure was around 2,000 CPU hours.

Availability

Source code for training new RGN models, as well as the trained RGN model used in this experiment, will be available shortly on GitHub.

1. AlQuraishi,M. (2018). End-to-end differentiable learning of protein structure. *bioRxiv*.
2. AlQuraishi,M. (2018) https://github.com/aqlaboratory/proteinnet, GitHub.

Protein structure refinement via short parallel molecular dynamics simulations

Yanjun Zhang and Sheng-You Huang*

Institute of Biophysics, School of Physics, Huazhong University of Science and Technology, Wuhan Hubei, 430074, People's Republic of China

huangsy@hust.edu.cn

Introduction

The three-dimensional structures of proteins are indispensable in the study of drug design and disease mechanism at the atom level. Unfortunately, there are still a lot of protein structures that have not been resolved because of experimental limitations. The development of computational protein structure predictions, especially for homology-modeling methods that can produce protein models with sufficient accuracy in some cases, gives us an alternative to obtain the structure of proteins that have not crystal structures. However, many predicted structures contain significant deficiencies such as incorrect loop conformations, secondary structures, and even domain orientations. Therefore, it is necessary to improve the quality of model structures via refinement methods. Since the protein structure refinement type was first introduced in the CASP9, the refinement methods have achieved great development, which can classify into two categories: MD-based sampling and iterative structure optimization[1-6]. The MD-based refinement methods were most successful because of the improvement of force fields and ensemble averaging approaches[1,4,6], but they need very time-consuming long simulations.

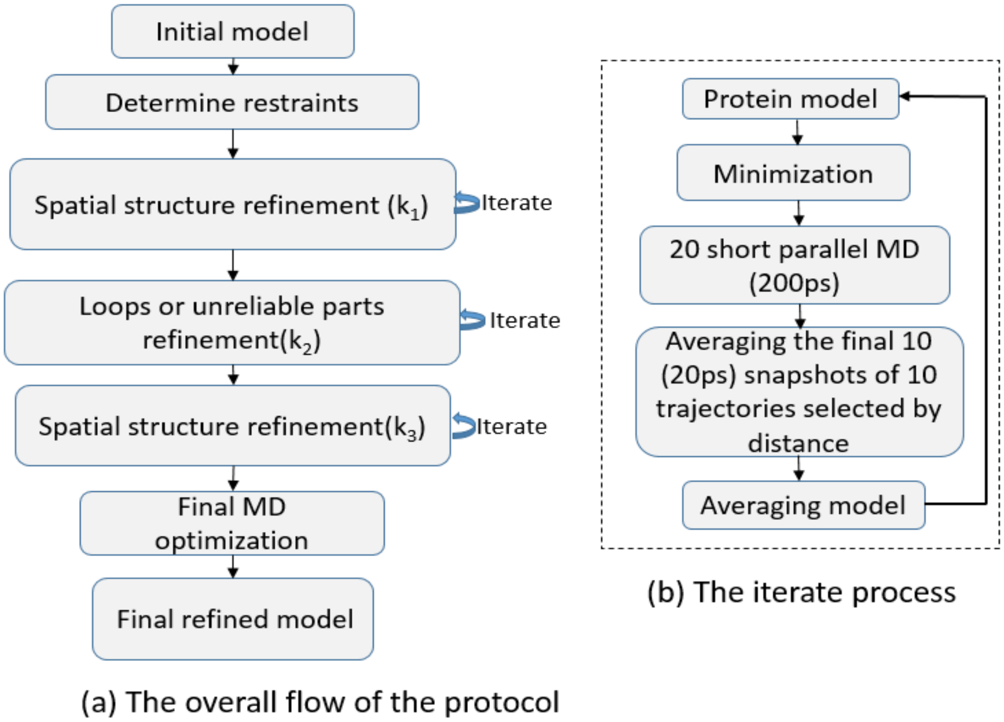
Here, we introduce a fast MD-based protocol for protein structure refinement. The goal of our protocol is to improve the protein structure quality by short parallel MD simulations within a few hours, which can greatly reduce time and computational resource. The results of our protocol on CASP12 are described in the following.

Method

Our protocol uses short parallel MD simulations to refine the protein structure so as to save time and computational resource. The refinement protocol is illustrated in Figure 1. For a given initial model, The LEAP module in AMBER was used to add the hydrogen atoms and missing heavy atoms. The structural restraints at different stages (say k_1, k_2, k_3) were determined by the flexibility and GDT-HA score of initial structure. Then, the spatial structure and loops or unreliable parts in the initial model were refined by iterative short parallel MD simulations [Figure 1 (a)]. We used the RMSF from a short MD simulation to determine the unreliable parts in the initial model. After a series of MD refinement, the snapshot with the lowest energy from the last-stage refinement was chosen as the final refined model.

Figure 1(b) shows the iterative short parallel MD simulation process. The iterations of short parallel MD simulations can greatly reduce the computational time of the refinement process, and usually finish a refinement process in a few hours. The AMBER14 ff14SB was used as the force-field parameters for all the simulations[6]. All the model structures were solvated in a cuboid periodic box with a cutoff 10 Å of TIP3P water modes. An appropriate number of Na^+ or Cl^- counterions were added to neutralize the system. A cutoff of 10 Å was chosen to calculate long-range electrostatics interactions. The Particle Mesh Ewald (PME) method was applied to calculate long-range electrostatic interactions. The SHAKE algorithm was used to constrain all the bonds involving hydrogen atoms.

Figure 1. A scheme of our protein structure refinement protocol.


Initial model
Determine restraints
Spatial structure refinement (k_1)
Iterate
Loops or unreliable parts refinement(k_2)
Iterate
Spatial structure refinement(k_3)
Iterate
Final MD optimization
Final refined model
(a) The overall flow of the protocol
Protein model
Minimization
20 short parallel MD (200ps)
Averaging the final 10 (20ps) snapshots of 10 trajectories selected by distance
Averaging model
(b) The iterate process

Results

We tested our protocol on the 24 CASP12 refinement targets with no more than 300 residues. It was shown that the average GDT-HA increase of model 1 is 1.42 points compared with the 1.3 points by the best-performed CASP12 groups. Among the 24 targets, eight models are improved by more than 3 points, and only two models become worse by 3 GDT-HA points, although the average improvement of RMSD value is very modest (-0.06). The protocol can refine a protein structure in 2~3 hours.

1. Hovan,L., Oleinikovas,V, Yalinca,H., Kryshtafovych,A., Saladino,G. & Gervasio,F.L. (2018) Assessment of the model refinement category in CASP12. *Proteins* **86**(S1), 152-167.
2. Park,H., Ovchinnikov,S., Kim,D,E., DiMaio,F. & Baker,D. (2018) Protein homology model refinement by large-scale energy optimization. *Proc. Natl. Acad. Sci. USA* ***115***, 3054-3059.
3. Feig,M. (2017) Computational protein structure refinement: almost there, yet still so far to go. *WIREs Comput. Mol. Sci.* **7**, e1307
4. Lee,G.R., Heo,L. & Seok,C. (2018) Simultaneous refinement of inaccurate local regions and overall structure in the CASP12 protein model refinement experiment. *Proteins* **86**(S1), 168-176.
5. Heo,L. & Feig,M. (2018) What makes it difficult to refine protein models further via molecular dynamics simulations? *Proteins* **86**(S1), 177-188.
6. Case,D., Berryman,J., Betz,R., Cerutti,D., Cheatham,I.T.E., Darden,T. et al. (2014) *AMBER 14*, University of California, San Francisco.

Use of Modified UNRES force field and replica-exchange molecular dynamics in physics-based template-free prediction of protein structures

Ł. Golon[1], M. Mozolewska[2], A.G. Lipska[1], A.K. Sieradzan[1*]

1 - Faculty of Chemistry, University of Gdańsk, Wita Stwosza 63, 80-308 Gdańsk, Poland, 2- - Institute of Computer Science, Polish Academy of Sciences, ul. Jana Kazimierza 5, Warsaw 01-248, Poland

adasko@sun1.chem.univ.gda.pl

Previous CASP experiments have shown that currently, physics-based approaches are less efficient than knowledge-based approaches in the prediction of proteins structure; however, their advantage is independence of structural databases. We used a coarse-grained force-field, as use of all-atom force fields is impractical in *de novo* simulations of protein structure due to excessive time and huge computer resources required. On the other hand, *de novo* simulations of even large proteins are feasible with coarse-grained force field that use highly reduced representation of polypeptide chains resulting in low resolution of acquired models.

In the last several years, we have been developing the physics-based united-residue (UNRES) force field for protein structure predictions and large-scale simulations of protein folding, together with a variety of methods for searching the conformational space[1]. Recently we introduced various improvements in UNRES. In the present CASP experiment we have tested an improved version of our physics-based coarse-grained force field.

Methods

In the UNRES model[1], a polypeptide chain is represented by a sequence of alpha-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 alpha-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 effective energy function depends on temperature[2]. We introduced a shielding function, which modifies the strength of the interactions between peptide-group dipoles depending on their screening from the solvent by side chains. The strength of the peptide-group − peptide-group interaction is linearly proportional to the volume of the first hydration sphere occupied by side-chains and we applied this term to maximum-likelihood optimized UNRES force field[3].

The structures of the target proteins were predicted by the following four-stage procedure. First, UNRES was employed to carry out Multiplexed Replica Exchange Molecular Dynamics (MREMD)[4] for target proteins. To speed up the search and improve accuracy, restraints were imposed on secondary structure based on secondary structure prediction by PSIPRED[5]. Those restraints were imposed both on torsional and valence bond angle. The strength of those restraints was proportional to PSIPRED score. Second, based on MREMD simulation results, Weighted-Histogram Analysis Method (WHAM) was used to calculate relative free energy of each structure of the last section of MREMD simulation[1]. Third, cluster analysis was employed to cluster the structures from an MREMD simulation. Five clusters with the lowest free energies were chosen as prediction candidates. Finally, in the fourth stage, the conformations closest to the respective average structures corresponding to the found clusters were converted to all-atom structures using the PULCHRA[6] and SCWRL[7] algorithms. These all-atom structures were submitted to the CASP website.

Results

We postpone the assessment of the approach until the official release of CASP13 results.

Availability

The UNRES package is available at 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, Chapter 8, pp. 107-122.
2. Liwo,A., Khalili,M., Czaplewski,C., Kalinowski,S., Ołdziej,S., Wachucik,K., Scheraga,H.A. (2007) Modification and optimization of the united-residue (UNRES) potential energy function for canonical simulations. I. Temperature dependence of the effective energy function and tests of the optimization method with single training proteins. *J. Phys. Chem. B* **111**, 260-285.
3. Krupa,P., Hałabis,A., Żmudzińska,W., Ołdziej,S., Scheraga,H.A., Liwo,A. (2017) Maximum likelihood calibration of the UNRES force field for simulation of protein structure and dynamics. *J. Chem. Inf. Model*. **57**, 2364-2377.
4. Sieradzan A.K (2015) Introduction of periodic boundary conditions into UNRES force field. J. Comput Chem, **36**, 940-946
5. 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. *J Chem. Theory Comput.* **5**, 627-640.
6. McGuffin,L.J., Bryson,K., Jones,D.T. (2000) The PSIPRED protein structure prediction server. *Bioinformatics*, **16**, 404-405.
7. Rotkiewicz,P., Skolnick,J. (2008) Fast procedure for reconstruction of full-atom protein models from reduced representations. *J. Comput. Chem.*, **29**, 1460-1465.
8. Wang,Q., Canutescu,A.A., Dunbrack,R.L. (2008) SCWRL and MolIDE: Computer Programs for Side-Chain Conformation Prediction and Homology Modeling. *Nat. Protoc. 3,* 1832-1847.

Fully Automated Prediction of Protein Tertiary Structures with Local Model Quality Scores Using the IntFOLD5 Server

L.J. McGuffin

School of Biological Sciences, University of Reading, Reading, UK

l.j.mcguffin@reading.ac.uk

The IntFOLD server[1] integrates our latest methods for: tertiary structure (TS) prediction, domain boundary prediction, prediction of intrinsically disordered regions, prediction of protein-ligand interactions and the global and local quality assessment (QA) of predicted 3D models of proteins. Following the success of the IntFOLD4 server at CASP12 [2,3], which used ModFOLD6_rank[4] to rank models, our primary focus for the IntFOLD5 server at CASP13 was the further improvement of global model ranking and local model quality scoring, using our newly improved ModFOLD7_rank method.

Methods

For CASP13, a bespoke version of the IntFOLD5 server was developed in order to return appropriately formatted results for just the tertiary structure (TS) prediction category. Additionally, the local quality assessment predictions (QMODE3) were returned as scores in the B-factor column of each TS model file. (Predictions in the QMODE1 & QMODE2 QA categories were also returned by our separate servers (see our ModFOLD7 and ModFOLDclust2 abstracts for details.)

Our 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 previous CASP successes in accurately predicting local errors in our models[5] by taking the global and local per-residue errors into consideration during the multiple template selection stage[6].

For the initial fold recognition stage, 14 different methods were installed and run in-house to generate up to 10 sequence-to-structure alignments each - resulting in up to 140 alternative single-template based models being generated for each CASP target. The following fold recognition methods were used: SP3 [7], SPARKS2 [7], HHsearch[8], COMA[9], SPARKSX[10], CNFsearch[11] and the 8 alternative threading methods that are integrated into the current LOMETS package[12] (PPA, dPPA, dPPA2, sPPA, MUSTER, wPPA, wdPPA and wMUSTER).

In the first stage of the IntFOLD5 TS method, all single-template models were assessed using ModFOLDclust2[13] in order to assign global and local model quality scores. Using the single template model quality scores, and other criteria involving template coverage, the corresponding alignments were then selected from each fold recognition method and used to build multiple-template models, with the aim of minimizing local errors in the final models. The alternative multi-template modelling alignment selection methods resulted in the generation of a new population of up to 124 multi-template models for each target. Additionally, I-TASSER *light*[14] (for sequence <500 residues; run in "light mode" with wall-time restricted to 5h) and HHpred[15] were used to generate 3 models each, which were then added to the final pool of alternative multi-template models for ranking. In the final stage of the method, the ~130 models in the final reference set were then evaluated using our new ModFOLD7_rank QA method and the top 5 ranked models were submitted as the final IntFOLD5 TS predictions (see our ModFOLD7 abstract from more details about our ModFOLD7_rank method).

Results

The IntFOLD5 server is continuously benchmarked using the CAMEO server[16] (identified as server 75). IntFOLD5 has been independently verified to be an improvement over our two previous methods

(IntFOLD3 & IntFOLD4). At the time of writing, IntFOLD5 ranks as the 2nd best 3D server according to the CAMEO lDDT scores based on pairwise comparisons (it is outperformed by only 1 public server in the benchmark).

Availability
The IntFOLD5 server is available at:
http://www.reading.ac.uk/bioinf/IntFOLD/IntFOLD5_form.html

1. McGuffin,L.J., Atkins,J., Salehe,B.R., Shuid,A.N. & Roche,D.B. (2015) IntFOLD: an integrated server for modelling protein structures and functions from amino acid sequences. *Nucleic Acids Res.* **43**, W169-73.
2. Kryshtafovych A., Monastyrskyy B., Fidelis K., Moult J., Schwede T. & Tramontano A. (2018) Evaluation of the template-based modeling in CASP12. *Proteins*. **86 S1**, 321-334.
3. McGuffin,L.J., Shuid,A.N., Kempster,R., Maghrabi,A.H.A., Nealon,J.O., Salehe,B.R., Atkins,J.D. & Roche,D.B. (2018) Accurate Template Based Modelling in CASP12 using the IntFOLD4-TS, ModFOLD6 and ReFOLD methods. *Proteins*., **86 S1**, 335-344.
4. Maghrabi,A.H.A. & McGuffin,L.J. (2017) ModFOLD6: an accurate web server for the global and local quality estimation of 3D models of proteins. *Nucleic Acids Res.* **45**, W416-W421.
5. McGuffin,L.J. & Roche,D.B. (2011) Automated tertiary structure prediction with accurate local model quality assessment using the IntFOLD-TS method. *Proteins*. **79 S10**, 137-46.
6. 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.
7. Zhou,H. & Zhou,Y. (2005) SPARKS2 and SP3 servers in CASP6. *Proteins*. **61 S7**, 152-156.
8. Söding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*. **21**, 951-96.
9. Margelevičius,M. & Venclovas,Č. (2010) Detection of distant evolutionary relationships between protein families using theory of sequence profile-profile comparisons. *BMC Bioinformatics*. **11**, 89.
10. 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 query and corresponding native properties of templates. *Bioinformatics*. **27**, 2076-2082.
11. Ma,J., Wang,S., Zhao,F. & Xu,J. (2013) Protein threading using context-specific alignment potential. *Bioinformatics*. **29**, i257-65.
12. Wu,S. & Zhang,Y. (2007) LOMETS: A local meta-threading-server for protein structure prediction. *Nucleic Acids Research*. **35**, 3375-3382.
13. 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.
14. Roy,A., Kucukural,A. & Zhang,Y. (2010) I-TASSER: a unified platform for automated protein structure and function prediction. *Nature Protocols*. **5**, 725-738.
15. Meier,A. & Söding,J. (2015) Automatic Prediction of Protein 3D Structures by Probabilistic Multi-template Homology Modeling. *PLoS Comput Biol*. **11**, e1004343.
16. Haas, J., Barbato, A., Behringer, D., Studer, G., Roth, S., Bertoni, M., Mostaguir, K., Gumienny, R. & Schwede, T. (2018) Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12. Proteins. **86 S1**, 387-398.

RigidQA: Protein Single-Model Quality Assessment based on rigidity analysis using a Support Vector Machine technique

Filip Jagodzinski[1*], and Renzhi Cao[2*]

1 - Department of Computer Science, Western Washington University, Bellingham, WA 98225, 2 - Department of Computer Science, Pacific Lutheran University, Tacoma, WA 98447

* Contributed equally: Filip.Jagodzinski@wwu.edu, caora@plu.edu

In CASP 13, we blindly tested our newly developed Single-Model Quality Assessment (QA) method RigidQA on targets in the Accuracy Estimation category. A Support Vector Machine (SVM) technique was used to train a machine learning model based on the features generated from rigidity analysis of protein decoys. Traditional QA methods [1–7] use the protein sequence properties and energies for quality evaluation of a protein structure, but none of them consider the rigidity properties of a biomolecule. Rigidity Analysis [8] is a combinatorial technique for identifying the rigid and flexible regions of proteins. It has been used to study thermostability properties of molecules [9], investigate protein cavities [10], and to infer the effects of amino acid substitutions on protein structure [11–13]. This is the first time that rigidity analysis has been used for protein structure prediction. Our SVM-based method relying on rigidity analysis is a new direction for tackling the QA problem, and aims to complement existing protein structure prediction approaches.

Methods

To train and test our model, we used a subset of the structures from CASP 9, 10, and 12 (63 targets from http://predictioncenter.org/download_area/) and 24 free modeling targets from the DeepQA datasets (https://www.cs.plu.edu/~caora//materials/softwares/DeepQA_cactus.tar.gz) [3]. We preprocessed and filtered the data for cases when the native structure of a target was not found, and finally we used 22,382 valid protein structure decoys for extracting the features. We used the KINARI software [14] and performed rigidity analysis on each protein structure decoy. In rigidity analysis, atoms and their chemical interactions are used to construct a mechanical model of a protein, a graph is constructed from the model, and pebble game algorithms [8] are used to analyze the rigidity of the associated graph. A total of 52 features were retained for each decoy, including biochemical metrics such as count of each type of amino acid, and rigidity features representing the count and type of each kind of rigid cluster. For example, a protein might have three rigid clusters made up of 16 atoms, three clusters made up for 20-30 atoms, etc. Cluster size tallies were recorded for clusters of size 2, 3, … 20, 21-30, 31-50, 51-100, 101-1000 atoms etc. All 52 features were normalized into the range between 0 and 1. We also divided all training data into bins with 0.2 step size based on real GDT-TS score (so there would be 5 different classes), and randomly selected the same number of data for each class. SVM was then used, and 5 cross validation was applied to avoid overfitting of the machine learning model.

Availability

The RigidQA software is available upon request.

1. Uziela, K., Menéndez Hurtado, D., Shu, N., Wallner, B. & Elofsson, A. (2017). ProQ3D: improved model quality assessments using deep learning. *Bioinformatics* **33**, 1578–1580
2. Cao, R. & Cheng, J. (2016). Protein single-model quality assessment by feature-based probability density functions. *Sci. Rep.* **6**, 23990
3. Cao, R., Bhattacharya, D., Hou, J. & Cheng, J. (2016). DeepQA: improving the estimation of single protein

model quality with deep belief networks. *BMC Bioinformatics* **17**, 495

4. Manavalan, B. & Lee, J. (2017). SVMQA: support-vector-machine-based protein single-model quality assessment. *Bioinformatics* **33**, 2496–2503
5. Olechnovič, K. & Venclovas, Č. (2017). VoroMQA: Assessment of protein structure quality using interatomic contact areas. *Proteins* **85**, 1131–1145
6. Derevyanko, G., Grudinin, S., Bengio, Y. & Lamoureux, G. (2018). Deep convolutional networks for quality assessment of protein folds. *Bioinformatics* doi:10.1093/bioinformatics/bty494
7. Kryshtafovych, A., Barbato, A., Monastyrskyy, B., Fidelis, K., Schwede, T. & Tramontano, A. (2016). Methods of model accuracy estimation can help selecting the best models from decoy sets: Assessment of model accuracy estimations in CASP11. *Proteins* **84 Suppl 1**, 349–369
8. Jacobs, D.J., Rader, A.J., Kuhn, L.A. & Thorpe, M.F. (2001). Protein flexibility predictions using graph theory. *Proteins* **44**, 150–165
9. Radestock, S. & Gohlke, H. (2008). Exploiting the Link between Protein Rigidity and Thermostability for Data-Driven Protein Engineering. *Eng. Life Sci.* **8**, 657–657
10. Mason, S., Chen, B.Y. & Jagodzinski, F. (2018). Exploring Protein Cavities through Rigidity Analysis. *Molecules* **23**,
11. Majeske, N. & Jagodzinski, F. (2018). Elucidating Which Pairwise Mutations Affect Protein Stability: An Exhaustive Big Data Approach. *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)* doi:10.1109/compsac.2018.00078
12. Dehghanpoor, R., Ricks, E., Hursh, K., Gunderson, S., Farhoodi, R., Haspel, N., Hutchinson, B. & Jagodzinski, F. (2018). Predicting the Effect of Single and Multiple Mutations on Protein Structural Stability. *Molecules* **23**,
13. Jagodzinski, F., Hardy, J. & Streinu, I. (2012). Using rigidity analysis to probe mutation-induced structural changes in proteins. *J. Bioinform. Comput. Biol.* **10**, 1242010
14. Fox, N., Jagodzinski, F., Li, Y. & Streinu, I. (2011). KINARI-Web: a server for protein rigidity analysis. *Nucleic Acids Res.* **39**, W177–83

DMPfold: a new deep learning-based method for protein tertiary structure prediction and model refinement

S.M. Kandathil[1,2], J.G. Greener[1,2] and D.T. Jones[1,2]

1 - University College London, Dept. of Computer Science, Gower Street, London WC1E 6BT, 2 - The Francis Crick Institute, 1 Midland Road, London NW1 1AT.

d.t.jones@ucl.ac.uk

Our entries in the TS and TR categories were based on a different implementation of our contact predictor, DeepMetaPSICOV or DMP (see abstract 'DMP') which was trained to reproduce inter-residue distance distributions. The only architectural change from the contact prediction version was that the final output layer was replaced by a 20-channel softmax layer so that a distance probability distribution could be estimated for each residue pair. The 20 distance bins used were of width 0.5 Å for distances < 8 Å, and 1 Å for distances >= 8 Å. The final bin represents all distances >= 19 Å.

Iterative Distance Map & Structure Generation (DMPfold)

DMPfold makes use of iterative generation and refinement of inter-residue distance distributions and hydrogen bond maps predicted using DeepMetaPSICOV (DMP). CNSsolve[1] is used to generate models from pseudo-NOE information derived from the DMP distance distributions, hydrogen bonding maps (main chain donor/acceptor pairs) and torsion angles.

In the first iteration, the contact maps and hydrogen bonding maps (asymmetric donor-acceptor pair maps) are predicted. Main chain dihedral angle inputs are generated with a deep convolutional network applied to the 60-dimensional MetaPSICOV inputs, where the final 2D output maps are projected down to 1D vectors using a spatial pooling operation. Upper and lower distance bounds, and predicted hydrogen bonds are converted into NOE-like constraints. For the Cβ-Cβ distances (Cα atoms for glycines), upper (60th percentile) and lower (10th percentile) bounds are estimated. H-bond H..O and N..O distances are constrained according to the distributions observed in high-resolution crystal structures.

Using the initial input constraints, 50 models are generated and clustered. The representative of the largest cluster taken, selected by estimated model accuracy using a combination of QMODCHECK[2] and MODELLER[3] DOPE-HR scores, and this model seeds the next iteration. The same DMP-distance and DMP-HB procedures described above are used, but with an additional input feature layer added, namely the Cα-Cα distance matrix for the seed structure. This allows new distances and H-bonds to be predicted using prior information of likely Cα-Cα distances from the previous iteration of 3D modelling. In this way, the combined contact prediction and structure generation procedure can evolve a better prediction at each iteration. Typically just 5 iterations are needed for convergence, and the whole process takes just one or two hours on a single workstation.

Although DMPfold was our primary method (94 models submitted), we also used FRAGFOLD[4] (34 models) and CONFOLD2[5] (20 models) to generate additional models (up to the maximum of 5) from DMP contact maps.

Refinement Category Predictions

The iterative DMP-distance predictor described above was used to predict Cβ-Cβ distances given the Cα distance matrix from each starting model. Different versions of the DMP ResNet were trained using different ranges of GDT-HA values for the starting models. This allowed us to condition the output of DMP-distance according to the starting GDT-HA (models with higher starting GDT-HA require less alteration). The predicted Cβ-Cβ distances were used to define distance restraints for MODELLER, using the 10th and 80th percentiles of the predicted distance distribution for each residue pair. In testing, we found that using the 5th lower bound percentile could sometimes produce improved results. Therefore, a second model was made using this lower bound. In both cases, we used MODELLER refinement to obtain a set of models, and the model with the lowest MODELLER score was submitted as the prediction. The per-residue confidence scores were obtained from the MODELLER decoys using the FILM3[6] MQA program (film3mqap), which uses the pairwise model RMSD in the ensemble to predict per-residue reliability. As the above protocol cannot currently be used with multimeric targets, for target R0979 we submitted the 5 lowest-scoring models obtained from 200 runs of Rosetta relax[7] in 'thorough' mode. In this case, the per-residue scores were the RMS fluctuation observed for each residue in the set of 200 models.

1. Brunger,A.T. (2007) Version 1.2 of the Crystallography and NMR System. *Nat. Protocols* **2**, 2728-33.
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. Pettitt,C.S., McGuffin,L.J. & Jones,D.T. (2005) Improving sequence-based fold recognition by using 3D model quality assessment. *Bioinformatics* **21**(17), 3509-15.
4. Adhikari,B. & Cheng,J. (2018) CONFOLD2: improved contact-driven ab initio protein structure modeling. *BMC Bioinformatics* **19**, 22.
5. 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.
6. Nugent,T. & Jones,D.T. (2012) Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. *Proc. Natl. Acad. Sci. U.S.A.* **109**(24):E1540-7.
7. Misura,K.M.S. & Baker,D. (2005) Progress and challenges in high-resolution refinement of protein structure models. *Proteins* **59**(1), 15-29.

Use of the coarse-grained UNRES force field in template-assisted and data-assisted prediction of protein structures

A.S. Karczyńska[1], K. Zięba[1], U. Uciechowska[1], M.A. Mozolewska[2], P. Krupa[3], E.A Lubecka1[4], A.G. Lipska[1], C. Sikorska[1], S.A. Samsonov[1], A.K. Sieradzan[1], A. Giełdoń[1], A. Liwo[1,5], R. Ślusarz[1], M. Ślusarz[1], J. Lee[5], K. Joo[5] and C. Czaplewski[1]

1 - Faculty of Chemistry, University of Gdańsk, Wita Stwosza 63, 80-308 Gdańsk, Poland, 2 - Institute of Computer Science, Polish Academy of Sciences, ul. Jana Kazimierza 5, Warsaw 01-248, Poland, 3 - Institute of Physics, Polish Academy of Sciences, Aleja Lotników 32/46, Warsaw, PL-02668, Poland, 4 - Institute of Informatics, Faculty of Mathematics, Physics, and Informatics, University of Gdańsk, Wita Stwosza 57, 80-308 Gdańsk, Poland, 5- Korea Institute for Advanced Study, 87 Hoegiro, Dongdaemun-gu, 130-722 Seoul, Republic of Korea,

cezary.czaplewski@ug.edu.pl

In the current CASP experiment, we tested a hybrid approach, which combines the physics-based coarse-grained United Residue (UNRES) force field[1] with knowledge-based information from templates (selected CASP-hosted server predictions). The fragments of a protein whose sequence is sufficiently similar to that of the proteins with known structures are modeled with knowledge-based method and the weakly similar parts with the physics-based method. Use of a coarse-grained force field is of advantage because of lower cost of energy and force evaluation and more extensive search of the conformational space.

Methods

In the UNRES model[1], a polypeptide chain is represented by a sequence of alpha-carbon atoms connected by virtual bonds with attached side chains. Only two interaction sites are used to represent each amino-acid residue; the united peptide groups (p) are located in the middle between two consecutive alpha-carbon atoms, and the united side chain group (SC). The UNRES potential-energy function was derived from the generalized cluster-cumulant expansion of a restricted free energy (RFE) function of polypeptide chains, which enable us to determine the functional forms of the multibody terms in UNRES. The effective energy function has been parameterized to reproduce structure and thermodynamics of selected training proteins[2].

In the first stage of the prediction procedure, top models from CASP-hosted server predictions (stage 2) are selected. Models from three servers that performed very well in the previous CASP exercises (BAKER-ROSETTA, Zhang-server, Quark) are supplemented by those highly ranked by DeepQA quality assessment[3]. Selected models are not used directly but are converted into the distance, the virtual angle, the virtual dihedral angle, and the sidechain positional restraints.[4] Models are compared to each other and restraints are imposed only on the fragments that have similar geometry for most of the models and those from completely diverse models are rejected.[4] In our earlier work[4-6] fragments were selected by visual inspection, while in this work we designed an automatic fragment selection procedure. In addition, pseudopotentials of the Dynamic Fragment Assembly (DFA) approach[7] were determined and added to the UNRES energy function.

For the refinement targets, the fragments corresponding to well-defined secondary structure and small uncertainty estimates (the B-factor columns of the template files) were selected from the templates provided and restraints were derived from these fragments. For the assisted-prediction targets, additional restraints were added, which were C(alpha)-C(alpha) distance restraints for crosslinking- and NMR-assisted targets and the experimental distance-distribution restraints for SAXS assisted targets.

In the second stage, the coarse-grained Multiplexed Replica Exchange Molecular Dynamics (MREMD)[8] simulations in the UNRES force field with restraint terms and DFA pseudopotentials, were run. In the next step, the Weighted-Histogram Analysis Method (WHAM) was used to calculate relative free energy of each structure of the last slice of the MREMD simulation[8]. Subsequently, a cluster analysis was used to obtain five clusters with the lowest free energies. The conformations closest to the respective average structures corresponding to the found clusters (cluster centroids) were converted to all-atom structures and refined by using restrained molecular dynamics simulations with the AMBER all-atom force field to obtain the models which were subsequently submitted.

Results

We postpone the assessment of the approach until the official release of CASP13 results.

Availability

The UNRES package is available at 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. ed. G. Voth, Taylor & Francis, Chapter 8, pp. 107-122.
2. Krupa,P., Hałabis,A., Żmudzińska,W., Ołdziej,S., Scheraga,H.A., Liwo,A. (2017) Maximum likelihood calibration of the UNRES force field for simulation of protein structure and dynamics. *J. Chem. Inf. Model.* **57**, 2364-2377.
3. Cao, R., Bhattacharya, D., Hou J., Cheng J. (2016) DeepQA: improving the estimation of single protein model quality with deep belief networks. BMC Bioinformatics **17**, 495.
4. Mozolewska, M., Krupa, P., Zaborowski, B., Liwo, A., Lee, J., Joo, K., Czaplewski, C. (2016) Use of Restraints from Consensus Fragments of Multiple Server Models To Enhance Protein-Structure Prediction Capability of the UNRES Force Field. *J. Chem. Inf. Model.* **56**, 2263-2279.
5. Karczyńska,A.S., Mozolewska,M.A., Krupa,P., Giełdoń,A., Liwo,A., Czaplewski,C. (2018) Prediction of protein structure with the coarse-grained UNRES force field assisted by small X-ray scattering data and knowledge-based information. *Proteins* **86 (S1)**, 228-239.
6. Karczyńska,A., Mozolewska,M.A., Krupa,P., Giełdoń,A., Bojarski,K.K., Zaborowski,B., Liwo,A., Ślusarz,R., Ślusarz,M., Lee,J., Joo,K., Czaplewski,C. (2018) Use of the UNRES force field in template-assisted prediction of protein structures and the refinement of server models: test with CASP12 targets, *J. Mol. Graph. Model.* **83**, 92-99.
7. Sasaki, T. N. & Sasai, M. (2004). A coarse-grained langevin molecular dynamics approach to protein structure reproduction. *Chemical Physics Letters* **402**, 102-106.
8. 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. *J Chem. Theory Comput.* **5**, 627-640.

Structure Prediction, Refinement, Quality Assessment, Contact Prediction, and Protein Docking in KiharaLab

Daisuke Kihara[1,2], Genki Terashi[1], Charles Christoffer[2], Woong-Hee Shin[1], Lyman Monroe[1], Tunde Aderinwale[2], and Sai Raghavendra Maddhuri Venkata Subramanya[2]

1 - Department of Biological Science, 2- Computer Science, Purdue University, West Lafayette, IN, USA

dkihara@purdue.edu

We submitted models in four categories of tertiary structure prediction to CASP13. They are regular (TS), refinement (TR), quality assessment (QA), contact prediction (RR). We also submitted protein docking models through CAPRI.

Methods

1. TS Regular targets for structure prediction

For monomeric targets, we selected 20 server models which were ranked by our QA method (below). All selected models were ranked by our QA method and manually inspected. Then, the selected models were relaxed by Rosetta-relax protocol[1]. For oligomeric targets, we searched for oligomer templates by HHsearch[2]. If appropriate template structures were not found, we used our protein-protein docking protocols, LZerD[3] and Multi-LZerD[4], with top-ranked server models.

2. TR Target Refinement

For refinement targets, we used our MD-based refinement protocol developed during the past CASP rounds[5]. The protocol uses an implicit solvent model FACTS with the CHARMM force field, and a dialectic constant of 2.0. (As described in another abstract under Kiharalab_RF2, we submitted TR models under another group name).

*3. QA Quality Assessmen*t

We used our new QA method that combined a new single-model QA method PRESCO2 and a machine-learning method. PRESCO2 searches similar residue environments observed in a query model in a reference database of representative native protein structures. The search results are subject to final quality prediction using machine learning method that was trained to distinguish near-native structures from other decoy structures. For the training datasets, we used QA models from CASP11 and CASP12.

4. RR Contact Prediction

For contact prediction category, we employed a consensus method of three different contact prediction programs, DeepContact[6], MetaPSICOV2[7], and CCMPred[8]. To predict a pairwise contact probability between residues, probabilities from three programs were weighted and summed. We tested the programs on CASP12 refinement targets and the weights of the programs were tuned based on the L/2 (L is a length of a target protein) accuracies. After the weighted sum, top 5L residue pairs were selected and the probabilities were re-scaled from 0.3 to 1.0.

5. Protein Docking

We submitted protein docking models through CAPRI. In principle, we followed our protocol reported for earlier rounds of CAPRI[9,10]. As described for TS above, we used template-based modeling and de novo docking with our LZerD suite. Decoys were ranked by the sum of the ranks of multiple scoring functions.

Acknowledgements

We are grateful for ITaP Research Computing at Purdue University for providing us the additional computational resource for this project. This work was partly supported by the National Institute of General Medical Sciences of the National Institutes of Health (R01GM123055) and the National Science Foundation (DMS161477).

1. Conway, P., Tyka, M. D., DiMaio, F., Konerding, D. E., & Baker, D. (2014). Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein Science*, **23**, 47-55.
2. Hildebrand, A., Remmert, M., Biegert, A., & Söding, J. (2009). Fast and accurate automatic structure prediction with HHpred. *Proteins: Structure, Function, and Bioinformatics*, **77**, 128-132.
3. Venkatraman, V., Yang, Y. D., Sael, L., & Kihara, D. (2009). Protein-protein docking using region-based 3D Zernike descriptors. *BMC bioinformatics*, **10**(1), 407.
4. Esquivel-Rodríguez, J., Yang, Y. D., & Kihara, D. (2012). Multi-LZerD: Multiple protein docking for asymmetric complexes. *Proteins: Structure, Function, and Bioinformatics*, **80**, 1818-1833.
5. Terashi, G. & Kihara, D. (2018). Protein structure model refinement in CASP12 using short and long molecular dynamics simulations in implicit solvent. *Proteins: Structure, Function, and Bioinformatics*, **86** Suppl 1, 189-201.
6. Liu, Y. Palmedo, P., Ye, Q. Berger, B., Peng, J. (2018) Enhancing Evolutionary Couplings with Deep Convolutional Neural Networks. *Cell Systems*, **6**:65-74.
7. Buchan, D. W. A. Jones, D. T. (2018) Improved protein contact predictions with the MetaPSICOV2 server in CASP12. *Proteins: Structure, Function, and Bioinformatics*, **86**, 78-83.
8. Seemayer, S. Gruber, M, Söding, J. (2014) CCMPred – fast and precise prediction of protein residue-residue contacts from correlated mutations. *Bioinformatics*, **30**, 3128-3130.
9. Peterson, L.X., Shin, W.H., Kim, H, & Kihara, D. (2018) Improved performance in CAPRI round 37 using LZerD docking and template-based modeling with combined scoring functions. *Proteins: Structure, Function, and Bioinformatics*, **86** Suppl 1, 311-320.
10. Peterson, L.X., Kim, H., Esquivel-Rodriguez, J., Roy, A, Han, X., Shin, W.H., Zhang, J., Terashi, G., Lee, M., & Kihara, D. (2017) Human and server docking prediction for CAPRI round 30-35 using LZerD with combined scoring functions. *Proteins: Structure, Function, and Bioinformatics*, **85**, 513-527.

Protein Structure Refinement using CABS and Contact Prediction

Daisuke Kihara[1,2], Woong-Hee Shin[1], Genki Terashi[1], and Sai Raghavendra Maddhuri Venkata Subramanya[2]

1- Department of Biological Science, 2- Computer Science, Purdue University, West Lafayette, IN, USA

Our lab has participated in the refinement category (TR) with two independent protocols. One uses molecular dynamics (MD) simulations[1], which refines details of model structures rather in a "conservative" fashion. This protocol is mentioned under the Kiharalab group. As a counterpart, we developed a new refinement protocol that uses the CABS model[2] in combination with residue-residue contact prediction. This is an adventurous method, which is aimed to perform larger conformational refinement. We have registered this adventurous method as a different human group, Kiharalab_RF2.

Methods

1. Initial structure preparation

To make diverse initial starting structures for the CABS-based refinement, an anisotropic network model by ProDy package[3] were used. CABS[2] is a coarse-grained protein structure model that uses a reduced structure representation and moves a model on a lattice using a Monte Carlo approach.

2. Restraints from contact prediction and PRESCO2

We put a contact prediction results as pairwise side-chain restraints in the CABS simulation. DeepContact[4], MetaPSICOV2[5], and CCMPred[6] were used for contact prediction. From each method, top L residue pairs were taken, where L is a sequence length of a refinement target. If residue pairs were selected more than two programs, then we put the pairs as side-chain restraints. We also took Cα-Cα restraints from PRESCO2 results, our new model quality assessment method, as follows: All server models of the refinement target were scored using PRESCO2 and top five models were selected. All pairwise Cα distances of the models were calculated, and if the standard deviation of the distances of a pair is less than 0.5 Å among the five models, then it is used as a Cα-Cα restraint.

3. Model selection

After CABS simulation, five models were selected considering four scoring functions, GOAP[7], RW+[8], dDFIRE[9], and OPUS-PSP[10]. The sum of the ranks by the four scoring functions. The colony energy[11] of each scoring function was also considered.

4. Side-chain remodeling and minimization

Side-chains of top five models were remodeled with SCWRL4[12]. The remodeled structures were further proceeded to short minimizations by CHARMM[13].

Acknowledgements

We are grateful for ITaP Research Computing at Purdue University for providing us the additional computational resource for this project. This work was partly supported by the National Institute of General Medical Sciences of the National Institutes of Health (R01GM123055) and the National Science Foundation (DMS161477).

1. Terashi, G. & Kihara, D. (2018). Protein structure model refinement in CASP12 using short and long molecular dynamics simulations in implicit solvent. *Proteins: Structure, Function, and Bioinformatics*, **86** Suppl 1, 189-201.

2. Kolinski, A. (2004) Protein modeling and structure prediction with a reduced representation. *Acta Biochimica Polonica*, **51**: 349-371.
3. Bakan, A. Meireles, L. M. Bahar, I. (2011) ProDy: Protein Dynamics Inferred from Theory and Experiments. *Bioinformatics*. **27**: 1575-1577.
4. Liu, Y. Palmedo, P., Ye, Q. Berger, B., Peng, J. (2018) Enhancing Evolutionary Couplings with Deep Convolutional Neural Networks. Cell Systems, 6:65-74.
5. Buchan, D. W. A. Jones, D. T. (2018) Improved protein contact predictions with the MetaPSICOV2 server in CASP12. *Proteins: Structure, Function, and Bioinformatics*, **86**:78-83.
6. Seemayer, S. Gruber, M, Söding, J. (2014) CCMPred – fast and precise prediction of protein residue-residue contacts from correlated mutations. *Bioinformatics*, **30**:3128-3130.
7. Zhou, J. Skolnick, H. (2011) GOAP: A Generalized Orientation-Dependent, All-Atom Statistical Potential for Protein Structure Prediction. *Biophysical Journal*, **101**: 2043-2052.
8. Zhang, J. Zhang. Y. (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.
9. 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.
10. Lu, M. Dousis, A. D. Ma, J. (2008) OPUS-PSP: an orientation-dependent statistical all-atom potential derived from side-chain packing. *Journal of Molecular Biology*, **376**: 288-301.
11. Lee, J. Seok, C. (2008) A statistical scoring scheme for protein-ligand docking: Consideration of entropic effect. *Proteins: Structure, Function, and Bioinformatics*, **70**: 1074-1083.
12. Krivov, G. G. Shapovalov, M. V. Dunbrack, R. L. (2009) Improved prediction of protein side-chain conformations with SCWRL4. *Proteins: Structure, Function, and Bioinformatics*, **77**: 778-795.
13. Brooks, B. R et al.,(2009) CHARMM: The Biomolecular simulation Program, *Journal of Computational Chemistry*, **30**: 1545-1615.

Constrained Docking using XL-MS and SAXS data

Dima Kozakov[1], Sandor Vajda[2,3], Kathryn Porter[2], Dzmitry Padhorny[1], Israel Desta[2], Dmitri Beglov[2], Mikhail Ignatov[1], Sergey Kotelnikov[1,4]

1- Laufer Center for Physical and Quantitative Biology, Stony Brook University; Departments of 2 Biomedical Engineering and 3 Chemistry, Boston University; 4- Moscow Institute of Physics and Technology

Fast Fourier Transform (FFT) based sampling enables global and systematic evaluation of molecular mechanics energy functions, expressed as sums of correlations, for all mutual orientations of two interacting proteins. In spite of the significant progress in scoring functions, docking generally improves when complemented by experimental data. Two most frequent types of data are from cross-linking and Small Angle X-ray Scattering (SAXS) experiments, and we have recently developed effective approaches to accounting for such information within the FFT based docking. In both cases the key to the method is the fact that FFT samples all possible configurations. Thus, given the results of cross-linking experiments we can simply select only those configurations that satisfy the data. To deal with SAXS data we have developed the generalized FFT-based algorithm FMFT-SAXS that performs massive SAXS computation on multiple conformations of the protein complex, exploiting the convolution-like form of the SAXS expression. In this round of CASP, we have applied those approaches to both cross-linking and SAXS data when such were available.

Methods

Model preparation.

For model preparation we either use the top template provided by HHPRED or, in difficult cases, build a "consensus" model for each target using the 150 server models provided by the CASP management committee. For each "easy" target most models had the same fold, with variations in loops and tails. Removal of the uncertain regions resulted in reliable "consensus" models that were used for docking.

Template based docking.

If a template of the biological complex is found then we model each monomer of the complex using Modeller, align separately to the template and co-minimize the resulting complex. Per rules of CAPRI we generate up to 10 models. If experimental data is available we filter the models using the data.

Free Docking.

Our free docking approach consists of two steps. The first step is running PIPER, a docking program that performs systematic search of complex conformations on a grid using the fast Fourier transform (FFT) correlation approach. The scoring function includes van der Waals interaction energy, an electrostatic energy term, and desolvation contributions calculated by a structure based pairwise potential. We can effectively account for cross-linking and SAXS data in the global search as described above.

The second step of the algorithm is clustering the top 1000 structures generated by PIPER using pairwise RMSD as the distance measure. The radius used in clustering is defined in terms of C_α interface RMSD. For each docked conformation we select the residues of the ligand that have any atom within 10 Å of any receptor atom, and calculate the C_α RMSD for these residues from the same residues in all other 999 ligands. Thus, clustering 1000 docked conformations involves computing a 1000 × 1000 matrix of pairwise C_α RMSD values. Based on the number of structures that a ligand has within a (default) cluster radius of 9 Å RMSD, we select the largest cluster and rank its cluster center as number one. The members of this cluster are removed from the matrix, and we select the next largest cluster and rank its center as number two, and so on. After clustering with this hierarchical approach, the ranked complexes are subjected to a straightforward (300 step and fixed backbone) minimization of the van der Waals energy using the CHARMM potential to remove potential side chain clashes.

Joint prediction of local and global protein model quality from atomic density distributions

G. Derevyanko[1], Y. Bengio[2] and G. Lamoureux[1,3]

1- Department of Chemistry and Biochemistry and Centre for Research in Molecular Modeling (CERMM), Concordia University, 2- Department of Computer Science and Operations Research, Université de Montréal, 3- Department of Chemistry and Center for Computational and Integrative Biology (CCIB), Rutgers University–Camden

georgy.derevyanko@gmail.com

Structural models of proteins can be assessed either globally, in terms of an overall score, or locally, in terms of the deviation of each residue from its position in the native structure. These two types of quality assessment (QA) methods usually have their own distinct algorithms, trained separately. In this work we present the 3DCNN-LQA algorithm, which was jointly trained to predict both the global and local quality of individual protein models.

Methods

The prediction algorithm relies on the 3D feature maps generated by our pre-trained 3DCNN model[1]. The 3DCNN model uses atomic density distributions as input, and is composed of four convolutional blocks that gradually reduce the spatial resolution of the data while increasing the number of features. For each protein residue (1 to L) along the sequence, local features are extracted from the 3D feature maps at the end of each of the first three blocks, from the grid cells corresponding to the position of the residue's C-alpha atom. This feature extraction procedure is similar to the one used in image captioning[2]. The L feature vectors are then fed to a bidirectional long short-term memory network (BiLSTM). At each position along the sequence, the forward and backward outputs of the BiLSTM are concatenated and fed to a single fully-connected layer with a sigmoidal nonlinearity. The output is treated as the local score prediction. The global score is predicted using the hidden states at the end of the forward and backward passes of the BiLSTM and the features from the last convolutional block of 3DCNN. These states/features are concatenated and fed to a 2-layer fully-connected neural network with ReLU and sigmoidal nonlinearities. The output of that network is treated as the global score prediction.

The model was trained on the CASP7 to CASP10 datasets, using the mean squared error on both the local and global scores as loss function. Decoys were preprocessed with SCWRL4[3] before being used for training or evaluation. The 3DCNN model used for feature extraction was not re-optimized during the training.

Results

Preliminary evaluation of our model was done using the CASP11 dataset. Performance of local quality predictions was measured using the per-decoy average correlation with lDDT score. For local QA, the model yields correlation of 0.42 for the CASP11 "stage 1" dataset and of 0.49 for the CASP11 "stage 2" dataset. In comparison ProQ4[4] yields a 0.56 per-decoy average correlation of local scores and lDDT for the full CASP11 dataset.

Performance of global quality predictions was evaluated using the per-target average correlation with GDT_TS score. For global QA, the model yields a GDT_TS correlation of 0.60 for the "stage 1" dataset and of 0.37 for the "stage 2" dataset. These results suggest that the representations learned by the 3DCNN model for the task of ranking protein decoys can be used for other tasks, such as local quality assessment.

Availability

The source code will be published in the github repository https://github.com/lamoureux-lab.

1. Derevyanko,G., Grudinin,S., Bengio,Y., & Lamoureux,G. (2018). Deep convolutional networks for quality assessment of protein folds, *Bioinformatics*, DOI:10.1093/bioinformatics/bty494.
2. Xu,K., Ba,J., Kiros,R., Cho,K., Courville,A., Salakhudinov,R., Zemel,R., & Bengio,Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. *Proceedings of the 32nd International Conference on Machine Learning,* PMLR 37:2048-2057.
3. Krivov,G.G., et al. (2009). Improved prediction of protein side-chain conformations with SCWRL4. *Proteins 77*, 778–795.
4. Menéndez Hurtado,D., et al. (2018). Deep transfer learning in the assessment of the quality of protein models. arXiv:1804.06281v1.

MELDxMD + homology: Physics based approach for protein folding and refinement

Alberto Perez[1,2], Cong Liu[1], Roy Nassar[1], James Robertson[1], Emiliano Brini[1] and Ken Dill[1]

1 Laufer Center, Stony Brook University, 2 Department of Chemistry, University of Florida

perez@chem.ufl.edu

MELD is an accelerator of Molecular Dynamics (MD) simualtions[1,2]. It accelerates force field atomistc simulations by using noisy, ambiguous and sparse data. However, for some proteins even with this advanced technique sampling remains an issue. Homology is fast and can provide good starting conformations for our MELDxMD approach. We feed in conformations from the predictions submitted by three servers: baker-rosettaserver,quark and Zhang-server. We use metagenomics[3,4] data when available and our general knowledge heuristics[2,5] for sampling has shown its ability to predict contacts form just sequence information. This information is noisy and ambiguous, suitable for MELD simulations. We have combined these different pipelines for sampling structures starting from sequence information. The replica exchange protocol allows both to identify what is the best server prediction that satisfies our information and force field and sample conformations that refine it.
We further used this method for refinement and NMR data.

Methods
We use the tleap program from the Amber[6] suite to generate a linear atomic structure based on the sequence and using the ff14SBside force field[7]. We used the MELD plugin to OpenMM[8] to carry out Hamiltonian and temperature replica exchange molecular dynamics[9] using the GbNeck2[10] implicit solvent. We used 30 replicas, running for at least 1µs.

We use the 15 predictions coming from the servers baker-rosettaserver, quark and Zhang-server. We minimize them with amber and seed each replica with a different prediction (each server prediction will be present twice in the replica ladder). In cases where we participated in refinement the protocol remains the same except we now have only 1 structure to seed the 30 replicas.

MELD revolves around the idea of incorporating information in the form of restraints and then asking for only a portion of that data to be satisfied. We make restraint groups out of individual restraints, setting an accuracy for the group. And put together groups of restraints into a collection. We have an accuracy number to satisfy for collections as well. Inside the replica exchange ladder the strength of the restraints is represented by αk. Where k is the force constant of the restraint and α is a scaling parameter that depends on replica. It is 0 at the highest replica and 1 (full force) at the lowest replica. α changes according to different functions that depend on the replica index. At each step of the simulation all restraints are evaluated, then sorted according to restraint energy and only the ones with the lowest energy up to the required accuracy (in group and collection) are enforced until the next step – where all evaluations happen again. In this way no information is ever lost and we always obey detailed balance – vital for a physics based methodology. Restraints follow our previous papers[2,5] and are explained below.

Coevolutionary data was produced from the gremlin[11] program using the HHsuite[12]. Each query sequence was iteratively searched against the May 2017 metaclust[13] database using jackhmmer [14], with bit score reporting and inclusion thresholds of 27, and maximum number of iterations set to 4. The output was filtered with hhfilter to include only 90% maximize pairwise sequence identity. GREMLIN residue-residue contacts that had at least 70% probability of being correct were used in MELD. Each residue-residue contact was treated as a MELD group of restraints between each pair of heavy atoms using as flat bottom harmonic restraints — 1 restraint had to be satisfied in each group. Those groups were inserted into a collection were X% of the groups were enforced.

We predicted secondary structure with psipred. We kept secondary structure for aminoacids that had secondary structure H or E and a score of 4 or higher. We impose the secondary structures as torsion and distance combination restraints. We dynamically enforce only 70% of this restraints as detailed below.

We place distance restraints on all pairs of hydrophobic residues ('ALA', 'ILE', 'LEU', 'MET', 'PHE', 'PRO', 'TRP', 'VAL'). For each pair of aminoacids we create pairwise restraints between all heavy atoms in the aminoacids, grouping them together so that only one has to be satisfied. The whole set of pairwise groups of restraints make a collection, and we ask that only 1.3*#Hydrophobic residues be satisfed. Each restraint is a flat-bottom harmonic potential. No energy penalty or force is applied beneath 5Å; quadratic up to 7Å and linear beyond.

We enforce distance restraints between residues in different predicted beta strands. N- -O pairs with flat bottom harmonic restraints parameters (flat: 3.5Å; quadratic up to 5Å and linear beyond). Only 0.45*#Beta strand residues need to be satisfied.

At the end of our simulations we cluster the lowest 5 replicas according to an average linkage procedure on alpha and beta carbons. We select the top 5 cluster centroid structures based on population. We then minimize each of these structures from the centroid structure to the cluster's average structure[15,16]. Additionally, in the case of refinement (TR targets) we performed a short 100ns Cα restrained simulation which we also clustered and submitted the top cluster along the other 4 clusters coming from the MELD simulation.

Results

The T1016 target was simulated starting from server predictions with MELDxMD hydrophobic, secondary structure, and metagenome-drive restraints. The best of the top five representative structures has a 4.6 Å backbone RMSD to the PDB reference. The MELDxMD results are in close agreement to the best server predictions, and much better than MELDxMD with the same restraints started from an extended conformation, rather than the server predictions. This suggests poor sampling when starting from extended, not surprising given the length of this 203 amino acid protein.

Availability

https://github.com/maccallumlab/meld.git

1. MacCallum, J. L., Perez, A. & Dill, K. Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 6985–6990 (2015).
2. Perez, A., MacCallum, J. L. & Dill, K. Accelerating molecular simulations of proteins using Bayesian inference on weak information. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 11846–11851 (2015).
3. Ovchinnikov, S., Kamisetty, H., Baker, D. & Roux, B. Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *Elife* **3,** e02030 (2014).
4. Ovchinnikov, S. *et al.* Protein structure determination using metagenome sequence data. *Science* **355,** 294–298 (2017).
5. Perez, A., Morrone, J. A., Brini, E., MacCallum, J. L. & Dill, K. Blind protein structure prediction using accelerated free-energy simulations. *Science Advances* **2,** e1601274–e1601274 (2016).
6. Case, D. A. *et al.* The Amber biomolecular simulation programs. *J. Comput. Chem.* **26,** 1668–1688 (2005).
7. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput* **11,** 3696–3713 (2015).
8. Eastman, P. *et al.* OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation. *J Chem Theory Comput* **9,** 461–469 (2013).
9. Sugita, Y. & Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters* **314,** 141–151 (1999).
10. Mongan, J., Simmerling, C., McCammon, J. A., Case, D. A. & Onufriev, A. Generalized Born model with a simple, robust molecular volume correction. *J Chem Theory Comput* **3,** 156–169 (2007).
11. Balakrishnan, S., Kamisetty, H., Carbonell, J. G., Lee, S.-I. & Langmead, C. J. Learning generative models for protein fold families. *Proteins* **79,** 1061–1078 (2011).
12. Söding, J., Biegert, A. & Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **33,** W244–8 (2005).

13. Steinegger, M. & Söding, J. Clustering huge protein sequence sets in linear time. *Nat Commun* **9,** 2542 (2018).
14. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7,** e1002195 (2011).
15. Feig, M. *et al.* Complete atomistic model of a bacterial cytoplasm for integrating physics, biochemistry, and systems biology. *Journal of Molecular Graphics and Modelling* **58,** 1–9 (2015).
16. Perez, A. *et al.* Extracting representative structures from protein conformational ensembles. *Proteins* **82,** 2671–2680 (2014).

MELDxMD Physics based approach for protein folding

Alberto Perez[1,2], Cong Liu[1], Roy Nassar[1], James Robertson[1], Emiliano Brini[1], Ken Dill[1]

1 Laufer Center, Stony Brook University, 2 Department of Chemistry, University of Florida

perez@chem.ufl.edu

Molecular Dynamics (MD) sampling and force fields have become more accurate, to the point that we can now fold structures of small proteins[1,2]. However, even proteins beyond 40 amino acids require too much sampling time. We have developed an accelerator of MD called MELD[3,4] (Modeling Employing Limited Data) that uses Bayesian inference to incorporate general knowledge (proteins have hydrophobic cores and beta strands pair up) and noisy secondary structure from psipred[5]. This knowledge is noisy, producing large amount of restraints for simulations – most of which are incorrect. Solving the problem of finding the folded structure and the sparse set of true restraints amongst all noisy restraints is much faster than either problem alone. In this way we have solved protein structures up to 110 residues. We used this method during CASP13 attempting to solve structures small enough for this method.

Methods

We use the tleap program from the Amber[6] suite to generate a linear atomic structure based on the sequence and using the ff14SBside force field[7]. We used the MELD plugin to OpenMM[8] to carry out Hamiltonian and temperature replica exchange molecular dynamics[9] using the GbNeck2[10] implicit solvent. We used 30 replicas, running for at least 1μs.

MELD revolves around the idea of incorporating information in the form of restraints and then asking for only a portion of that data to be satisfied. We make restraint groups out of individual restraints, setting an accuracy for the group. And put together groups of restraints into a collection. We have an accuracy number to satisfy for collections as well. Inside the replica exchange ladder the strength of the restraints is represented by αk. Where k is the force constant of the restraint and α is a scaling parameter that depends on replica. It is 0 at the highest replica and 1 (full force) at the lowest replica. α changes according to different functions that depend on the replica index. At each step of the simulation all restraints are evaluated, then sorted according to restraint energy and only the ones with the lowest energy up to the required accuracy (in group and collection) are enforced until the next step – where all evaluations happen again. In this way no information is ever lost and we always obey detailed balance – vital for a physics based methodology. Restraints follow our previous papers[4,11] and are explained below.

We predicted secondary structure with psipred. We kept secondary structure for aminoacids that had secondary structure H or E and a score of 4 or higher. We impose the secondary structures as torsion and distance combination restraints. We dynamically enforce only 70% of this restraints as detailed below.

We place distance restraints on all pairs of hydrophobic residues ('ALA', 'ILE', 'LEU', 'MET', 'PHE', 'PRO', 'TRP', 'VAL'). For each pair of aminoacids we create pairwise restraints between all heavy atoms in the aminoacids, grouping them together so that only one has to be satisfied. The whole set of pairwise groups of restraints make a collection, and we ask that only 1.3*#Hydrophobic residues be satisfed. Each restraint is a flat-bottom harmonic potential. No energy penalty or force is applied beneath 5Å; quadratic up to 7Å and linear beyond.

We enforce distance restraints between residues in different predicted beta strands. N- -O pairs with flat bottom harmonic restraints parameters (flat: 3.5Å; quadratic up to 5Å and linear beyond). Only 0.45*#Beta strand residues need to be satisfied.

At the end of our simulations we cluster the lowest 5 replicas according to an average linkage procedure on alpha and beta carbons. We select the top 5 cluster centroid structures based on population. We then minimize each of these structures from the centroid structure to the cluster's average structure[12,13].

Results

Most of the PDBs for CASP targets have not been released yet. For the ones that have: For target T0958 our prediction has 3.8 Å RMSD to the 6BTC PDB reference. And T0955 has sub-1.0 Å RMSD to the 5W9F PDB reference.

Separate from CASP we found that this method folds 20/41 nonthreadable[14] proteins less than 100 amino acids to within 4.0 Å RMSD. We found that we were mostly limited by the force field we used and to poor sampling in beta-rich proteins.

Availability

https://github.com/maccallumlab/meld.git

1. Lindorff-Larsen, K., Piana, S., Dror, R. O. & Shaw, D. E. How fast-folding proteins fold. *Science* **334,** 517–520 (2011).
2. Nguyen, H., Maier, J., Huang, H., Perrone, V. & Simmerling, C. Folding simulations for proteins with diverse topologies are accessible in days with a physics-based force field and implicit solvent. *J. Am. Chem. Soc.* **136,** 13959–13962 (2014).
3. MacCallum, J. L., Perez, A. & Dill, K. Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 6985–6990 (2015).
4. Perez, A., MacCallum, J. L. & Dill, K. Accelerating molecular simulations of proteins using Bayesian inference on weak information. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 11846–11851 (2015).
5. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292,** 195–202 (1999).
6. Case, D. A. *et al.* The Amber biomolecular simulation programs. *J. Comput. Chem.* **26,** 1668–1688 (2005).
7. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput* **11,** 3696–3713 (2015).
8. Eastman, P. *et al.* OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation. *J Chem Theory Comput* **9,** 461–469 (2013).
9. Sugita, Y. & Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters* **314,** 141–151 (1999).
10. Mongan, J., Simmerling, C., McCammon, J. A., Case, D. A. & Onufriev, A. Generalized Born model with a simple, robust molecular volume correction. *J Chem Theory Comput* **3,** 156–169 (2007).
11. Perez, A., Morrone, J. A., Brini, E., MacCallum, J. L. & Dill, K. Blind protein structure prediction using accelerated free-energy simulations. *Science Advances* **2,** e1601274–e1601274 (2016).
12. Feig, M. *et al.* Complete atomistic model of a bacterial cytoplasm for integrating physics, biochemistry, and systems biology. *Journal of Molecular Graphics and Modelling* **58,** 1–9 (2015).
13. Perez, A. *et al.* Extracting representative structures from protein conformational ensembles. *Proteins* **82,** 2671–2680 (2014).
14. Skolnick, J. & Zhou, H. Why Is There a Glass Ceiling for Threading Based Protein Structure Prediction Methods? *J Phys Chem B* acs.jpcb.6b09517 (2016). doi:10.1021/acs.jpcb.6b09517

MELDxMD + Metagenomic data: Physics based approach for protein folding

Alberto Perez[1,2], Cong Liu[1], Roy Nassar[1], James Robertson[1], Emiliano Brini[1], Ken Dill[1]

1 Laufer Center, Stony Brook University, 2 Department of Chemistry, University of Florida

perez@chem.ufl.edu

MELD is an accelerator of Molecular Dynamics (MD) simualtions[1,2]. It accelerates force field atomistcc simulations by using noisy, ambiguous and sparse data. In its most basic form we can frame general knowledge (e.g. proteins have hydrophobic cores) to fold proteins up to 110 residues long[2,3]. Recently metagenomics has shown its ability to predict contacts form just sequence information[4,5]. This information is noisy and ambiguous, suitable for MELD simulations. We have combined these two protocols for sampling structures starting from sequence information.

Methods

We use the tleap program from the Amber[6] suite to generate a linear atomic structure based on the sequence and using the ff14SBside force field[7]. We used the MELD plugin to OpenMM[8] to carry out Hamiltonian and temperature replica exchange molecular dynamics[9] using the GbNeck2[10] implicit solvent. We used 30 replicas, running for at least 1μs.

MELD revolves around the idea of incorporating information in the form of restraints and then asking for only a portion of that data to be satisfied. We make restraint groups out of individual restraints, setting an accuracy for the group. And put together groups of restraints into a collection. We have an accuracy number to satisfy for collections as well. Inside the replica exchange ladder the strength of the restraints is represented by αk. Where k is the force constant of the restraint and α is a scaling parameter that depends on replica. It is 0 at the highest replica and 1 (full force) at the lowest replica. α changes according to different functions that depend on the replica index. At each step of the simulation all restraints are evaluated, then sorted according to restraint energy and only the ones with the lowest energy up to the required accuracy (in group and collection) are enforced until the next step – where all evaluations happen again. In this way no information is ever lost and we always obey detailed balance – vital for a physics based methodology. Restraints follow our previous papers[2,3] and are explained below.

Coevolutionary data was produced from the GREMLIN [11] program using the HHsuite[12]. Each query sequence was iteratively searched against the May 2017 metaclust[13] database using jackhmmer [14], with bit score reporting and inclusion thresholds of 27, and maximum number of iterations set to 4. The output was filtered with hhfilter to include only 90% maximize pairwise sequence identity. GREMLIN residue-residue contacts that had at least 70% probability of being correct were used in MELD. Each residue-residue contact was treated as a MELD group of restraints between each pair of heavy atoms using as flat bottom harmonic restraints — 1 restraint had to be satisfied in each group. Those groups were inserted into a collection were X% of the groups were enforced.

We predicted secondary structure with psipred. We kept secondary structure for aminoacids that had secondary structure H or E and a score of 4 or higher. We impose the secondary structures as torsion and distance combination restraints. We dynamically enforce only 70% of this restraints as detailed below.

We place distance restraints on all pairs of hydrophobic residues ('ALA', 'ILE', 'LEU', 'MET', 'PHE', 'PRO', 'TRP', 'VAL'). For each pair of aminoacids we create pairwise restraints between all heavy atoms in the aminoacids, grouping them together so that only one has to be satisfied. The whole set of pairwise groups of restraints make a collection, and we ask that only 1.3*#Hydrophobic residues be satisfed. Each restraint is a flat-bottom harmonic potential. No energy penalty or force is applied beneath 5Å; quadratic up to 7Å and linear beyond.

We enforce distance restraints between residues in different predicted beta strands. N- -O pairs with flat bottom harmonic restraints parameters (flat: 3.5Å; quadratic up to 5Å and linear beyond). Only 0.45*#Beta strand residues need to be satisfied.

At the end of our simulations we cluster the lowest 5 replicas according to an average linkage procedure on alpha and beta carbons. We select the top 5 cluster centroid structures based on population. We then minimize each of these structures from the centroid structure to the cluster's average structure[15,16].

Results

We are awaiting results to determine if we metagenome-derived coevolutionary contacts improve our protein folding routine.

Availability

https://github.com/maccallumlab/meld.git

1. MacCallum, J. L., Perez, A. & Dill, K. Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 6985–6990 (2015).
2. Perez, A., MacCallum, J. L. & Dill, K. Accelerating molecular simulations of proteins using Bayesian inference on weak information. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 11846–11851 (2015).
3. Perez, A., Morrone, J. A., Brini, E., MacCallum, J. L. & Dill, K. Blind protein structure prediction using accelerated free-energy simulations. *Science Advances* **2,** e1601274–e1601274 (2016).
4. Ovchinnikov, S., Kamisetty, H., Baker, D. & Roux, B. Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *Elife* **3,** e02030 (2014).
5. Ovchinnikov, S. *et al.* Protein structure determination using metagenome sequence data. *Science* **355,** 294–298 (2017).
6. Case, D. A. *et al.* The Amber biomolecular simulation programs. *J. Comput. Chem.* **26,** 1668–1688 (2005).
7. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput* **11,** 3696–3713 (2015).
8. Eastman, P. *et al.* OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation. *J Chem Theory Comput* **9,** 461–469 (2013).
9. Sugita, Y. & Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters* **314,** 141–151 (1999).
10. Mongan, J., Simmerling, C., McCammon, J. A., Case, D. A. & Onufriev, A. Generalized Born model with a simple, robust molecular volume correction. *J Chem Theory Comput* **3,** 156–169 (2007).
11. Balakrishnan, S., Kamisetty, H., Carbonell, J. G., Lee, S.-I. & Langmead, C. J. Learning generative models for protein fold families. *Proteins* **79,** 1061–1078 (2011).
12. Söding, J., Biegert, A. & Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **33,** W244–8 (2005).
13. Steinegger, M. & Söding, J. Clustering huge protein sequence sets in linear time. *Nat Commun* **9,** 2542 (2018).
14. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7,** e1002195 (2011).
15. Feig, M. *et al.* Complete atomistic model of a bacterial cytoplasm for integrating physics, biochemistry, and systems biology. *Journal of Molecular Graphics and Modelling* **58,** 1–9 (2015).
16. Perez, A. *et al.* Extracting representative structures from protein conformational ensembles. *Proteins* **82,** 2671–2680 (2014).

Single-model quality assessment of individual protein models using random forest and statistical potentials

Tong Liu, Zheng Wang

Department of Computer Science, University of Miami

Zheng.Wang@miami.edu

We developed two single-model prediction servers (i.e., MASS1 and MASS2) for quality assessment (QA) of individual protein models, which were originated from our previous four servers developed and benchmarked in CASP11 [1]. We widely used multiple new features from protein potentials, including self-defined or self-implemented protein statistical potentials and Rosetta energies. Particularly, we designed six different types of protein statistical potentials including pseudo-bond angle potential (PAP), accessible surface potential at the atomic level (ASPA), sequence separation-dependent potential (SSDP), contact-dependent potential (CDP), relative solvent accessibility potential (RSAP), and volume-dependent potential (VDP). We redesigned or re-implemented torsion angle potential (TAP), centrosymmetric burial potential (CSP), accessible surface potential at the residue level (ASPR), and distance-dependent potential (DDP).

Methods

For each single protein model, we generated 70 features, part of which are the features we used in our previous work [1-3]. The 70 features we used in CASP13 experiments can be categorized into seven classes: (1) consistency of predicted and assigned secondary structure (i.e., Q3 and SOV scores [4]) and solvent accessibility; (2) three existing statistical potential energies of protein models, including RWplus, GOAP, and DRIRE; (3) pseudo amino acid composition of protein sequences; (4) radius of gyration; (5) residue-residue contact information; (6) newly designed or modified protein statistical potentials; (7) Rosetta energies. For each single residue of a protein model, we extracted 64 features out of the 70 features without including classes of (2), (4), and (5).

For residue-specific deviation predictions, we trained two Random Forest models (MASS1 and MASS2) with the number of trees equal to 1500 and 2500, respectively. For global score predictions from MASS1, we trained a global prediction server (i.e., Random Forest model) using the 70 features. For global score predictions from MASS2, we combined all predicted residue deviations into a global score. The training and cross-validation data sets were extracted from CASP 9 and CASP 10. The data sets for blind test were obtained from CASP 11 and CASP 12.

Availability

The two QA servers (MASS1 and MASS2) in CASP 13 experiments are available at http://dna.cs.miami.edu/MASS/QA.html.

1. Liu, T., Wang, Y., Eickholt, J. & Wang, Z. Benchmarking Deep Networks for Predicting Residue-Specific Quality of Individual Protein Models in CASP11. Scientific reports 6, 19301 (2016).
2. Wang, Z., Eickholt, J. & Cheng, J. APOLLO: a quality assessment service for single and multiple protein models. Bioinformatics 27, 1715-1716 (2011).
3. Cao, R., Wang, Z., Wang, Y. & Cheng, J. SMOQ: a tool for predicting the absolute residue-specific quality of a single protein model with support vector machines. BMC bioinformatics. 15, 120 (2014).
4. Liu, T. & Wang, Z. SOV_refine: A further refined definition of segment overlap score and its significance for protein structure similarity. Source Code for Biology and Medicine 13, 1, doi:10.1186/s13029-018-0068-7 (2018).

Contact map prediction with DeepPPP and template-free 3D structure prediction with SSThread2.0

K.J. Maurice

kevin_maurice@hotmail.com

Contact maps are predicted by DeepPPP (Deep learning for Protein Property Prediction) (unpublished) using correlated mutations and deep neural networks. SSThread[1] first predicts the structure of contacting α-helices and β-strands (secondary structure elements [SSEs]). Then overlapping pair predictions are assembled to create a set of core structure predictions. The loops and side chain conformations are then predicted.

Methods

DeepPPP predicts five structural properties: 8-state secondary structure, half-sphere exposure[2], real-value backbone torsion angles, contact maps and domain boundaries. The first three properties are predicted using three separate networks. A Hidden Markov Model (HMM) obtained from searching with HHblits[3] against the UniClust30[4] database is used. The inputs to the three networks are the sequence and the amino acid & transmission frequencies from the HMM. The networks have embedding, convolutional and recurrent layers. The inputs for contact map prediction include the amino acid and transmission frequencies from the HMM, the predicted secondary structure, the predicted half-sphere exposures, the predicted torsion angles, the sequence distance between residue pairs, the mutual information at each residue pair and the direct mutual information calculated by CCMpred[5]. The network uses convolutional layers. Domain boundary prediction uses alignments to three databases: UniClust30[4], Pfam[6] and SCOP[7]. Amino acid and transmission frequencies from the HMM are also used. The network uses convolutional and recurrent layers. Domain boundary prediction is designed to use template alignments when available but does not require them.

SSThread was run on individual domains using domain boundaries predicted by DeepPPP. A database of contacting SSE pairs was created by clustering pairs taken from a set of experimental structures. For each protein, many pair predictions are made using the pair database and the protein sequence. Then an ensemble of core predictions is generated using a non-stochastic greedy search in which structures containing an increasing number of SSEs are created by merging two smaller structures that can joined by an overlapping pair prediction. To predict loops, segments are taken from a set of experimental structures that have similar end orientations to the gap and are high scoring. The segments are then closed using Cyclic Coordinate Descent[8].

Predictions are scored using a knowledge-based potential (KBP) and using predictions from DeepPPP. The KBP terms include orientation-dependent residue to residue contacts, half-sphere exposures, backbone torsion angles, compactness and SSE lengths. The KBP uses sequences that are homologous to the protein sequence obtained from the search against the UniClust30 database. During pair prediction an additional score is used to force the distribution of predictions among the residues of the protein to accurately reflect the secondary structure prediction.

All atom predictions are generated by predicting side chain conformations with SIDEpro[9] followed by a brief energy minimization with GROMACS[10] using the AMBER[11] force field. The predictions are then clustered by RMSD to reduce redundancy. The all-atom KBP dDFIRE[12] is used to select the top 10 predictions. Refinement is carried out by GalaxyRefine[13]. The 5 submitted structures were selected from the top 10 by manual inspection, preferring native-like structures.

Availability

DeepPPP and SSThread are available as stand-alone programs free for non-commercial use at www.kjmaurice.com/downloads.html.

1. Maurice,K.J. (2014) SSThread: Template-free protein structure prediction by threading pairs of contacting secondary structures followed by assembly of overlapping pairs. *J. Comput. Chem.* **35**, 644-656.
2. Hamelryck,T. (2005) An amino acid has two sides: a new 2D measure provides a different view of solvent exposure. *Proteins* **59**, 38-48.
3. Remmert,M., Biegert,A., Hauser,A. & Soding,J. (2011) HHblits: lightening-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **25**, 173-175.
4. Mirdita,M., von den Driesch,L., Galiez,C., Martin,M.J., Soding,J. & Steinegger,M. (2017) Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res.* **45**, D170-D176.
5. Seemayer,S., Gruber,M. & Soding,J. (2014) CCMpred—fast and precise prediction of protein residue-residue contacts from correlated mutations. *Bioinformatics* **30**, 3128-3130.
6. Finn,R.D., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Mistry,J., Mitchell,A.L., Potter,S.C., Punta,M., Qureshi,M., Sangrador-Vegas,A., Salazar,G.A., Tate,J. & Bateman,A. (2016) The Pfam families database: towards a more sustainable future. *Nucleic Acids Res.* **44**, D279-D285.
7. Fox,N.K., Brenner,S.E. & Chandonia,J.M. (2014) SCOPe: Structural Classification of Proteins-extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.* **42**, D304-D309.
8. Canutescu,A.A. & Dunbrack,R.L. Jr. (2003) Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Sci.* **12**, 963-972.
9. Nagata,K., Randall,A. & Baldi,P. (2012) SIDEpro: a novel machine learning approach for the fast and accurate prediction of side-chain conformations. *Proteins* **80**, 142-153.
10. Berendsen,H.J.C, van der Spoel,R. & van Drunen,R. (1995) GROMACS: A message-passing parallel molecular dynamics implementation. *Comp. Phys. Comm.* **91**, 43-56.
11. Wang,J., Cieplak,P. & Kollman,P.A. (2000) How Well Does a Restrained Electrostatic Potential (RESP) Model Perform in Calculating Conformational Energies of Organic and Biological Molecules? *J. Comp. Chem.* **21**, 1049-1074.
12. Yang,Y. & Zhou,Y. (2008) Specific interactions for ab initio folding of protein terminal regions with secondary structures. *Proteins* **72**, 793-803.
13. Heo,L., Park,H. & Seok,C. (2013) GalaxyRefine: Protein structure refinement driven by side-chain repacking. *Nucleic Acids Res.* **41**, W384-W388.

Manual Prediction of Protein Tertiary and Quaternary Structures and 3D Model Refinement

L.J. McGuffin[1], R. Adiyaman[1], D.A. Brakenridge[1], J.O. Nealon[1], L.S. Philomina[1] and A.N. Shuid[1,2]

1 - School of Biological Sciences, University of Reading, Reading, UK, 2 - Infectomics Cluster, Advanced Medical and Dental Institute, University of Science Malaysia, Pulau Pinang, Malaysia

l.j.mcguffin@reading.ac.uk

For our manual predictions we used several components from our latest servers[1,2,3] (also see our IntFOLD5 and ModFOLD7 server abstracts). For our tertiary structure (TS) predictions we made use of the CASP hosted 3D server models, which we ranked using ModFOLD7_rank and then refined with the our new refinement method (ReFOLD2). For our quaternary structure predictions, we used a docking and template based approach (MultiFOLD) along with our newly developed quality assessment method (ModFOLDdock). Finally, clues from likely ligand binding sites (predicted with FunFOLD3), aided our manual evaluation of submitted models.

Methods

Tertiary structure predictions: The server models were ranked according the ModFOLD7_rank global quality scores (see our ModFOLD7 abstract). The top ranked initial model was then selected and submitted to the ReFOLD2 and MultiFOLD pipelines described below. For each model, the ModFOLD7 predicted per-residue error scores were added into the B-factor column for each set of atom records.

Refinement (ReFOLD2): For the refinement of 3D models of proteins we used a modified version of our automated ReFOLD method[3]. Our new refinement pipeline, ReFOLD2, consisted of three protocols that were similar to the original version. The first protocol used a rapid iterative strategy (i3Drefine[4]) and the second employed a more CPU/GPU intensive molecular dynamic simulation strategy (using NAMD[5]) to refine each starting model.

The major new step for ReFOLD version 2 was the modification of the second protocol, which included the introduction of molecular dynamics simulations that were guided by the per-residue accuracy scores obtained from ModFOLD7. The per-residue accuracy scores were used to identify the poorly predicted regions, which were then targeted for refinement to improve the overall model quality. A new restraint strategy was applied by putting a threshold based on the per-residue accuracy scores (either 2, 3 or 5 Å) during the molecular dynamic simulation. For each starting model, the threshold was determined by considering the distribution of the per-residue accuracy scores. Refined models generated from the first two protocols were then assessed and ranked using ModFOLD7_rank. The third protocol was a combination of the first 2 approaches, where the top ranked model from the 2nd protocol was then further refined using i3Drefine. Finally, all of the refined models generated by each of these protocols and the starting model were pooled and re-ranked again using ModFOLD7_rank and the final top 5 models were selected and submitted.

Quaternary structure predictions (MultiFOLD): The highest scoring models from the ReFOLD2 procedure, described above, were used to generate predicted quaternary structures using LZerD[6], MEGADOCK[7], FRODOCK[8], PatchDock[9] and ZDOCK[10] for dimeric complexes, and M-ZDOCK[11] and Multi-LZerD[12] for multimeric complexes. In addition to the docking strategy, a multimeric fold recognition approach was also deployed. The fold template lists (with PDB and chain IDs) generated by the IntFOLD server[1] were filtered using multimeric data extracted from PISA[13] for each template. Model assemblies were then constructed using TM-align[14] for structural superposition of tertiary models onto assemblies and PyMOL was used for visualization and manual quality checking of the template generated models. The final predicted quaternary structures were then ranked for submission using the newly developed ModFOLDdock method described below. Furthermore, the information from our FunFOLD3 method (regarding the function and locations of putative bound ligands) along with visual inspection was used for some targets in order to manually filter the modelled complexes.

Quaternary structure model quality assessment (ModFOLDdock): The ModFOLDdock protocol uses a hybrid consensus approach for producing both global and local (interface residue) scores for predicted quaternary structures. The ModFOLDdock global score was taken as the mean score from four individual methods: ProQDock[15], QSscoreJury, DockQJury and ModFOLDIA. For each interacting pair of chains in a modelled complex, the ProQDock scores were simply taken and averaged to produce a global score for the complete assembly. For the QSscoreJury and DockQJury methods, pairwise comparisons were made for each quaternary structure model to every other model made for the target and then the mean QS[16] and DockQ[17] scores were calculated. The ModFOLDIA method also carries out structure based comparisons of alternative oligomer models and can produce both global and local/per-residue interface scores. The first stage of the ModFOLDIA method was to identify the interface residues in the model to be scored (defined as <= 5Å between the heavy atoms in different chains) and then obtain the minimum contact distance (D_{min}) for each contacting residue. The second stage was to locate the equivalent residues in all other models and then obtain the mean minimum distances of those residues in all other models ($MeanD_{min}$). The final IA score for each of the interface residues i in the model was the absolute difference in the S_i from the mean S_i : $IA = 1\text{-}|S_i\text{-}MeanS_i|$, where $S_i = 1/(1+(D_{min}/20)2)$ and $MeanS_i = 1/(1+(MeanD_{min}/20)2)$. The global ModFOLDIA score for a model was then taken as the total interface score (sum of residue scores) normalized by the maximum of either the number of residues in the interface or the mean number of interface residues across all models for the same target.

Availability
Our software will be freely available after publication from:
http://www.reading.ac.uk/bioinf/downloads/
Server methods are available via:
http://www.reading.ac.uk/bioinf/

1. McGuffin,L.J., Atkins,J., Salehe,B.R., Shuid,A.N. & Roche,D.B. (2015) IntFOLD: an integrated server for modelling protein structures and functions from amino acid sequences. *Nucleic Acids Res.* **43**, W169-73.
2. Maghrabi,A.H.A. & McGuffin,L.J. (2017) ModFOLD6: an accurate web server for the global and local quality estimation of 3D models of proteins. *Nucleic Acids Res.* **45**, W416-W421.
3. Shuid,A.N., Kempster,R. & McGuffin,L.J. (2017) ReFOLD: a server for the refinement of 3D protein models guided by accurate quality estimates. *Nucleic Acids Res.* **45**, W422-W428.
4. Bhattacharya,D. & Cheng,J. (2013) i3Drefine software for protein 3D structure refinement and its assessment in CASP10. *PLoS One*. **8**, e69648.
5. Phillips,J.C., Braun,R., Wang,W., Gumbart,J., Tajkhorshid,E., Villa,E., Chipot,C., Skeel,R.D., Kalé,L. & Schulten,K.J. (2005) Scalable molecular dynamics with NAMD. *Comput. Chem.* **26**, 1781-802.
6. Venkatraman,V., Yang,Y.D., Sael,L. & Kihara,D. (2009) Protein-protein docking using region-based 3D Zernike descriptors. *BMC Bioinformatics*. **10**, 407.
7. Ohue,M., Shimoda,T., Suzuki,S., Matsuzaki,Y., Ishida,T. & Akiyama,Y. (2014) MEGADOCK 4.0: an ultra–high-performance protein–protein docking software for heterogeneous supercomputers. *Bioinformatics*. **30**, 3281–3283.
8. Garzon,J.I., Lopéz-Blanco,J.R., Pons,C., Kovacs,J., Abagyan,R., Fernandez-Recio,J. & Chacon,P. (2009) FRODOCK: a new approach for fast rotational protein–protein docking. *Bioinformatics*. **25**, 2544–2551.
9. Duhovny,D., Nussinov,R. & Wolfson,H.J. (2002) Efficient unbound docking of rigid molecules, in: International Workshop on Algorithms in Bioinformatics. Springer, pp. 185–200.
10. Chen,R., Li,L. & Weng,Z. (2003) ZDOCK: An initial-stage protein-docking algorithm. *Proteins*. **52**, 80–87.
11. Pierce,B., Tong,W. & Weng,Z. (2005) M-ZDOCK: a grid-based approach for Cn symmetric multimer docking. *Bioinformatics*. **21**, 1472–1478.

12. Esquivel-Rodríguez,J., Yang,Y.D. & Kihara,D. (2012) Multi-LZerD: Multiple protein docking for asymmetric complexes. *Proteins*. **80**, 1818-1833.
13. Krissinel,E. & Henrick,K. (2007) Inference of Macromolecular Assemblies from Crystalline State. *J. Mol. Biol.* **372**, 774–797.
14. Zhang,Y. & Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302-9.
15. Basu,S. & Wallner,B. (2016) Finding correct protein-protein docking models using ProQDock. Bioinformatics. **32**, i262-i270.
16. Bertoni,M., Kiefer,F., Biasini,M., Bordoli,L. & Schwede,T. (2017) Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. Sci Rep. **7**, 10480.
17. Basu,S. & Wallner,B. (2016) DockQ: A Quality Measure for Protein-Protein Docking Models. PLoS One. **11**, e0161879.

The human MESHI group in CASP13

T. Sidi and C. Keasar

Ben Gurion University of the Negev

keasar@bgu.ac.il

The human MESHI group submitted tertiary structure predictions based on CASP server models. Our major aim was to explore the power of EMA beyond the limits of the EMA track. Most importantly, to apply EMA at the domains level rather than whole chains. As for apparent single domain targets, we tried to improve selections of MESHI-enrich-server (See the MESHI-enrich-server abstract). To this end, we generated hundreds of close variants of the top scoring server models. These variations were then scored, often leading to re-ranking. In many cases though, due to time constraints (both human and computer) we simply submitted the energy-minimized server models with the highest MESHI-enrich-server ranks.

In addition, the human MESHI group also submitted EMA predictions, most often simply resending MESHI-enrich-score predictions. In some cases, however, visual inspection led to the use of MESHI-corr-score instead.

Methods

Multi-domain targets: Multi domain targets and domain boundaries were identified by superposition, and visual inspection of the top ranking decoys. Then, MESHI-enrich-score was applied to each domain separately. The top scoring domain models, typically from different server decoys, were combined by MEDELLER[1], energy minimized by MESHI OPTIMIZER[2] and submitted. When the top scoring server decoys did not agree on domain boundaries, we repeated this procedure with different domain splitting and chose which complete models to submit by MESHI-enrich-score.

Apparent single-domain targets: Time permitting, we supplied the top ranking server models as templates to MEDELLER[1], which generated a few hundreds of slightly different copies for each of them. These new decoys were then ranked by MESHI-enrich-score.

We used structural consistency among the best models as an estimate of local quality (temperature factor). To this end, we structurally aligned each of the five submitted decoys with the other 19 top scoring decoys. The average distance between a C-alpha atom and its counterparts served as the quality estimate of all the residue's atoms.

1. Sali A. & Blundell T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* 234, 779-815
2. Kalisman N., Levy A., Maximova T., Zafriri-Lynn S., Reshef D. & Keasar C. (2005) *MESHI*: a new library of Java classes for molecular modeling. *Bioinformatics 21:3931-3932.*

Loss-function effect on EMA performance

T. Sidi and C. Keasar

Ben Gurion University of the Negev

keasar@bgu.ac.il

In CASP13 we aimed to study the effect of loss-functions, which guide training, on the performance of EMA methods that use machine learning. To this end we created three EMA servers, which were practically clones, differing only in the loss-function that was used during their training. Overall, the computational scheme is very similar to the one used in CASP12[1]. In a nutshell, the servers implement MESHI-Score, an ensemble learning method, whose individual predictors (5000 for each server) are trained using Monte-Carlo Simulated-Annealing optimization.

Methods

Data: Decoys are represented by vectors of features extracted after sidechain repacking[2] and restrained energy minimization[3]. Overall we consider 129 features including in-house energy terms[4,5], non-bonded energy terms[6-8], and compatibility with predicted secondary-structure and solvent accessability[9,10]. None of the features corresponds to the similarity of a decoy to another one.

Training database: The individual predictors were trained on a non-redundant (at the target level) dataset of server decoys from CASP9-12. Overall 305 targets and 73605 decoys.

Individual predictors: Individual predictors are pairs of non-linear functions of the features[1]: a score, estimating the decoy's accuracy, and a weight that indicates the score's reliability. The parameters of the score are learned by Monte-Carlo Simulated Annealing optimization of a loss function. In CASP13 we tested three loss-functions, which are the medians (over all the targets) of three target-specific functions:

(I) loss-enrichment – for each target the loss is proportional to the inverse of the fraction of the 10% top quality decoys (in terms of gdt_ts) within the 10% top scoring decoys.

(II) loss-correlation – for each target the loss is the opposite of Spearman's correlation between observed and predicted quality.

(III) loss-contacts – for each target the loss is the opposite of the Matthews correlation coefficient between the observed and predicted contact maps (where a contact is a distance of 8Å or less between C-alpha atoms.

Score-functions: MESHI scores are the weighted medians of the scores generated by ensembles of individual predictors. Each of the three servers: MESHI-enrich-server, MESHI-corr-server and MESHI-server employed an ensemble of 5000 individual predictors, which were independently trained on loss-enrichment, loss-correlation and loss-contacts respectively.

Human intervention: Overall, the prediction pipeline was automatic, and most targets did not need any human intervention. Yet, in some of the larger targets a few decoys failed the preprocessing minimization step. These were typically decoys with long extended and unstructured segments. In order to comply with the CASP forms we manually assigned these decoys arbitrary low quality estimates.

Availability

The current version of the MESHI-package including the score functions is available in available in https://github.com/meshiprot/meshi/releases.

1. Elofsson E., Joo K., Keasar C., Lee J., Maghrabi A.H.A, Manavalan B., McGuffin L.J., Ménendez D.H., Mirabello C., Pilstål R., Sidi T., Uziela K., & Wallner B. (2018) Methods for estimation of model accuracy in CASP12. Proteins 86:361–373
2. Krivov G.G., Shapovalov M.V. & Dunbrack R.K. (2009) Improved prediction of protein side-chain conformations with SCWRL4. Proteins, 77, 778–795.
3. Kalisman N., Levy A., Maximova T., Zafriri-Lynn S., Reshef D. & Keasar C. (2005) MESHI: a new library of Java classes for molecular modeling. Bioinformatics 21:3931-3932.
4. Amir E.D., Kalisman N. & Keasar C. (2008) Differentiable, multi-dimensional, knowledge-based energy terms for torsion angle probabilities and propensities. Proteins, 72, 62–73.
5. Levy-Moonshine A., Amir E.D. & Keasar C. (2009) Enhancement of beta-sheet assembly by cooperative hydrogen bonds potential. Bioinformatics, 25, 2639–2645.
6. Zhou H. and Skolnick J. (2011) GOAP: A Generalized Orientation-Dependent, All-Atom Statistical Potential for Protein Structure Prediction. Biophysical Journal, 101, 2043–2052.
7. Summa C.M. and Levitt M. (2007) Near-native structure refinement using in vacuo energy minimization. PNAS, 104, 3177–3182.
8. Samudrala R. and Moult J. (1998) An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction1. Journal of Molecular Biology, 275, 895–916.
9. McGuffin L.J., Bryson K. & Jones D.J. (2000) The PSIPRED protein structure prediction server. Bioinformatics 16:404-405
10. Wang S. Peng J., Ma J. & Xu J. (2016) Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. Scientific Reports volume 6, Article number: 18962

The MESHI-server pipeline for template-based structure prediction

T. Sidi and C. Keasar

Ben Gurion University of the Negev

keasar@bgu.ac.il

MESHI server took part in two CASP13 tracks: EMA and tertiary structure predictions. The EMA predictions are discussed in the MESHI-enrich-server abstract. Here we describe the tertiary structure prediction.

Methods

Our aim in building the server was to test the contribution of EMA to the prediction quality of a simple server that uses established methods. Specifically, MESHI-server used HHPRED[1] for template identification and alignment and MODELLER[2] for building the models. We aimed to generate up to 2,000 models per target and pick the top five using MESHI-enrich-server (see the EMA servers abstract). In practice for quite a few targets, most notably the larger ones, we had to make do with a lower number of decoys (down to a few dozen) due to CPU limitations.

We used structural consistency among the best models as an estimate of local quality (temperature factor). To this end, we structurally aligned each of the five submitted decoys with the other 19 top scoring decoys. The average distance between a C-alpha atom and its counterparts served as the quality estimate of all the residue's atoms.

Availability

The current version of the MESHI-package including the score functions is available in available in https://github.com/meshiprot/meshi/releases.

1. Zimmermann L., Stephens A., Nam S.Z., Rau D., Kübler J., Lozajic M., Gabler F., Söding J, Lupas A.N., Alva V. (2018) A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *J. Mol. Biol.* S0022-2836(17)30587-9
2. Sali A. & Blundell T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* 234, 779-815

Automated 3D Model Quality Assessment using the ModFOLD7 Server

L.J. McGuffin and A.H.A. Maghrabi

School of Biological Sciences, University of Reading, Reading, UK

l.j.mcguffin@reading.ac.uk

The ModFOLD7 server is the latest version of our web resource for the Quality Assessment (QA) of 3D models of proteins[1,2,3].

Methods

ModFOLD7 is our new approach to QA that combines the strengths of multiple pure-single and quasi-single model methods for improving prediction accuracy. For CASP13, our emphasis was on increasing the accuracy of per-residue assessments for single models, single model ranking and score consistency. Each model was considered individually using six pure-single model methods: CDA[3], SSA[3], ProQ2[4], ProQ2D[5], ProQ3D[5] and VoroMQA[6]. Additionally, sets of reference 3D models generated using IntFOLD5 (see our other abstract) were used to score models using four alternative quasi-single model methods: DBA[3], MF5s[3], MFcQs[3] and ResQ[7]. Neural networks (NNs) were then used to combine the component per-residue/local quality scores from each of the 10 alternative scoring methods, resulting in a final consensus of per-residue quality scores for each model.

***Component per-residue/local quality scoring methods**:*
The ModFOLD7 NNs were trained using two separate target functions for each residue in a model: the superposition based S-score used previously[3] and the residue contact based lDDT score[8]. For the method trained using the lDDT score (ModFOLD7_res_lddt), the per-residue similarity scores were calculated using a simple multilayer perceptron (MLP). The MLP input consisted of a sliding window (size=5) of per-residue scores from all 10 of the methods described above, and the output was a single quality score for each residue in the model (50 inputs, 25 hidden, 1 output). For the method straind using the S-score (ModFOLD7_res), the per-residue similarity scores were also calculated using an MLP with a sliding window (size=5) of per-residue scores, but this time only 7 of the 10 methods were used as inputs - all apart from the ProQ2, CDA and SSA scores (therefore 35 inputs, 18 hidden, 1 output). The RSNNS package for R was used to construct the NNs, which were trained using data derived from the evaluation of CASP11 & 12 server models versus native structures. For both of the per-residue scoring methods, the similarity scores, s, for each residue were converted back to distances, d, with $d = 3.5\sqrt{((1/s) -1)}$.

Global scoring methods:
Global scores were calculated by taking the mean per-residue scores (the sum of the per-residue similarity scores divided by sequence lengths) for each of the 10 individual component methods, described above, plus the NN output from ModFOLD7_res and ModFOLD7_res_lddt. Furthermore, 3 additional quasi-single global model quality scores were generated for each model based on the original ModFOLDclust, ModFOLDclustQ and ModFOLDclust2 global scoring methods[9] (in a similar vein to the ModFOLD4_single and ModFOLD5_single global scores, tested in CASP10 and CASP11 respectively). Thus, we ended up with 15 alternative global QA scores, which could be combined in various ways in order to optimize for the different facets of the quality estimation problem. We registered three ModFOLD7 global scoring variants:

The ModFOLD7 global score (the mean per-residue NN output score from ModFOLD7_res) considered alone was found to have a good balance of performance both for correlations of predicted versus observed scores and rankings of the top models.

The ModFOLD7_cor global score variant *((MFcQs + DBA + ProQ3D + ResQ + ModFOLD7_res)/5)* was found to be an optimal combination for producing good correlations with the observed scores, i.e. the predicted global quality scores produced should produce closer to linear correlations with the observed global quality scores.

The ModFOLD7_rank global score variant *((CDA + SSA + VoroMQA + ModFOLD7_res + ModFOLD7res_lDDT)/5)* was found to be an optimal combination for ranking, i.e. the top ranked models (top 1) should be closer to the highest accuracy, but the relationship between predicted and observed scores may not be linear.

The local scores of the ModFOLD7 and ModFOLD_rank variants used the output from the ModFOLD7_res NN, whereas the ModFOLD_cor variant used the local scores from the ModFOLD7_res_lddt NN.

Results

The ModFOLD7 server is continuously benchmarked in the Model Quality Estimation (QE) category using the CAMEO server[10] (identified as server 28). The method has been independently verified to be an improvement on our previous leading ModFOLD4 & ModFOLD6 methods. At the time of writing ModFOLD7 ranks among the top few QE servers.

Availability

The ModFOLD7 server is available at:
http://www.reading.ac.uk/bioinf/ModFOLD/ModFOLD7_form.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., Buenavista,M.T. & Roche,D.B. (2013) The ModFOLD4 Server for the Quality Assessment of 3D Protein Models. *Nucleic Acids Res.* **41**, W368-72.
3. Maghrabi,A.H.A. & McGuffin,L.J. (2017) ModFOLD6: an accurate web server for the global and local quality estimation of 3D models of proteins. *Nucleic Acids Res.* **45**, W416-W421.
4. Uziela,K. & Wallner,B. (2016) ProQ2: estimation of model accuracy implemented in Rosetta. *Bioinformatics*. **32**, 1411-3.
5. Uziela,K., Menéndez Hurtado,D., Shu,N., Wallner,B. & Elofsson,A. (2017) ProQ3D: improved model quality assessments using deep learning. *Bioinformatics*. **33**, 1578-1580.
6. Olechnovič,K. & Venclovas,Č. (2017) VoroMQA: Assessment of protein structure quality using interatomic contact areas. *Proteins*. **85**, 1131-1145.
7. Yang,J., Wang,Y. & Zhang,Y. (2016) ResQ: An Approach to Unified Estimation of B-Factor and Residue-Specific Error in Protein Structure Prediction. *J Mol Biol*. **428**, 693-701.
8. 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.
9. Mariani,V., Biasini,M., Barbato,A. & Schwede,T. (2013) lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*. **29**, 2722-8.
10. Haas,J., Barbato,A., Behringer,D., Studer,G., Roth,S., Bertoni,M., Mostaguir,K., Gumienny,R. & Schwede,T. (2018) Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12. *Proteins*. **86 S1**, 387-398.

Automated 3D Model Quality Assessment using ModFOLDclust2

L.J. McGuffin

School of Biological Sciences, University of Reading, Reading, UK

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 server tested during CASP13 was identical to that tested during the CASP9, CASP10, CASP11 & CASP12 experiments. The ModFOLDclust2 method was originally developed to provide increased prediction accuracy, over the original ModFOLDclust method[3,4], 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. 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 simply combining the per-residue scores with those from ModFOLDclustQ.

Results

ModFOLDclust2 has been independently evaluated by the CASP assessors since CASP9 and has consistently ranked among the top performing QA methods[2,7,8,9].

Availability

ModFOLDclust2 can be run as an 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 S10**, 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 S9**, 185-190.
5. Ben-David,M., Noivirt-Brik,O., Paz,A., Prilusky,J., Sussman,J.L. & Levy,Y. (2009) Assessment of CASP8 structure predictions for template free targets. *Proteins*. **77 S9**, 50-65.
6. Zhang,Y. & Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins*. **57**, 702-710.
7. Kryshtafovych,A., Barbato,A., Fidelis,K., Monastyrskyy,B., Schwede,T., & Tramontano,A. (2013) Assessment of the assessment: evaluation of the model quality estimates in CASP10. *Proteins*. **82 S2**, 112-26.
8. Kryshtafovych,A., Barbato,A., Monastyrskyy,B., Fidelis,K., Schwede,T., & Tramontano,A. (2015) Methods of model accuracy estimation can help selecting the best models from decoy sets: Assessment of model accuracy estimations in CASP11. *Proteins*. **84 S1**, 349-69.
9. Kryshtafovych,A., Monastyrskyy,B., Fidelis,K., Schwede, T, & Tramontano,A. (2018) Assessment of model accuracy estimations in CASP12. *Proteins*. **86 S1**, 345-360.

MUfoldQA_M and MUfoldQA_T, New Consensus-based Protein Model QA Methods

Wenbo Wang[1], Dong Xu[1,2], and Yi Shang[1]

1-Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, Missouri, 65211, USA,
2 - Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, Missouri, 65211, USA

wwr34@mail.missouri.edu

MUfoldQA_M and MUfoldQA_T are two new consensus-based protein model QA methods we deployed in CASP13. They apply new techniques to select and score reference models, and new techniques to use reference models to score candidate models in a consensus fashion. Significant improvement has been achieved over the naïve consensus method.

Methods

Both methods take a target protein sequence and a set of candidate models to be scored as input and return the scores of the candidate models in the range of 0 and 1, with 1 being identical to the native structure of the protein.

The basic idea of both methods is to use a set of reference models to score each candidate models. Their difference is in how the reference models are selected and how the score of a candidate model are calculate given a set of reference models. Incorporated with our previously developed quasi-single-model QA, both methods first obtain a set of templates by searching the PDB database with the target protein sequence. Next, a subset of the candidate models is selected as reference and each reference model is scored based on its similarity to the templates, to calculate a global score and local scores for all positions of the sequence. Finally, each candidate model is scored based on its similarity to the reference models, weighted by the scores of the reference models. Specifically, the process of the methods is as follows:

Step 1. Obtain a set of templates. The target protein sequence is used to search the PDB database with Blast[1] and HHsearch[2], respectively, to find similar proteins as templates. The templates are sorted based on a comprehensive consideration of their E-value, percentage of identical sequences, and coverage. Then, a certain number（K) of top templates are retained from Blast and HHsearch results, respectively. The number K for Blast and HHsearch results can be different and is determined separately as follows: if the top 10 templates cover all C-alpha positions of the target sequence, then K is set to 10; otherwise, increase K by including more top templates that could contribute to combined coverage until all C-alpha positions of the target sequence are covered by some templates.

Step 2. Select a subset of the candidate models as reference models. The two methods use different algorithms to select the reference models.

MUfoldQA_T:

a) Calculate a score for each candidate model and classify this target into one of different categories. Each category has a set of pre-determined parameters for the bandpass filter in the next step. If the size of the candidate model set is smaller than 50, use MUfoldQA_S, a quasi-single-model QA method we developed for CASP12, to calculate the score. Then, the average score of all candidate models is used to classify this target into one of 3 categories: easy, medium and hard, based on pre-determined, fixed thresholds. Otherwise, MQAPRank[3] is used to calculate the score. Then, the average score of all candidate models is used to classify this target into one of 4 categories: easy, easy-medium, medium-hard, and hard, based on pre-determined, fixed thresholds.
b) Initially, assume all models are reference models. Sort the models based on the scores computed in step a). Starting from the model with the highest score, compare it with all other models. The models that are either very similar to or very different from the current

model are removed from the reference model pool based on the bandpass filter parameters. Then, move on to the remaining model with next highest score. Repeat this step until all remaining models have been performed this test upon and the final remaining models are the reference models.

MUfoldQA_M:

If the size of the candidate model set is smaller than 50, use the entire set as the reference model set. Otherwise, sort all candidate models using their MQAPRank scores and choose the top 45% as reference model set A, and the top 83% as reference model set B.

Step 3. Use MUfoldQA_S to calculate the local scores (weights), **W**, for each C-alpha position of each reference model using the templates generated in Step 1. MUfoldQA_T has a single **W** based on its single reference model set, whereas MUfoldQA_M has two, $\mathbf{W}^{\mathbf{A}}$ and $\mathbf{W}^{\mathbf{B}}$, based on its two reference model sets, respectively.

Step 4. Calculate GDT-TS, **G**, between each candidate model and each reference model.

Step 5. Calculate the final scores for candidate models.

MUfoldQA_T: the final score of a candidate model is a weighted-sum of **W** and **G.**

MUfoldQA_M: the final score of a candidate model is a linear combination of the weighted-sum of $\mathbf{W}^{\mathbf{A}}$ and **G**, and the weighted-sum of $\mathbf{W}^{\mathbf{B}}$ and **G**.

Results

Both methods have been tested on CASP12 targets for both CASP stage 1 QA task (72 targets each with up to 20 models) and stage 2 QA task (72 targets each with up to 150 models). Their results of Pearson correlation and average GDT-TS difference are compared with those of the naïve consensus method in the following table. For stage 1 QA task, the two new methods outperform naïve consensus by more than 25% on Pearson correlation and 20% on average GDT-TS difference. They are also significantly better than naïve consensus on the stage 2 QA task.

	Stage 1 QA task		Stage 2 QA task	
	Pearson Correlation	Avg. GDT-TS Diff	Pearson Correlation	Avg. GDT-TS Diff
Naïve Consensus	0.64340	0.05126	0.77897	0.06305
MUfoldQA_M	0.80687	0.04091	0.83159	0.05276
MUfoldQA_T	0.80506	0.04001	0.84826	0.05022

1. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
2. Soding,J. (2004). Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.
3. Jing,X. Dong,Q. Liu,X & Liu,B. (2015). Protein model quality assessment by learning-to-rank. *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, 91-96.

Tertiary Structure Prediction Assisted by SAXS data, Probabilistic Modeling, Deep Learning, and Contact Predictions

Jie Hou and Jianlin Cheng

Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65211, USA

chengji@missouri.edu

Small Angle X-ray Scattering (SAXS) is an experimental technique that has the potential to generate low-resolution data regarding the overall shape of proteins to aid protein modeling. MULTICOM applied the probabilistic sampling and simulated annealing optimization, which was guided by an energy function with the restraints from SAXS data as energy terms, to generate structural conformations matching SAXS profiles. The entire prediction process has four major steps: (1) manual quality inspection of SAXS data, (2) domain prediction of target sequence, (3) SAXS-guided model generation/domain-assembly, and (4) automated SAXS-assisted model quality assessment. The final selected models were submitted to the SAXS-assisted structure modeling category of CASP13.

Methods

All the CASP13 server models of a SAXS-assisted target and its experimental SAXS data were firstly collected. The preliminary analysis of small-angle scattering data was conducted, including (1) monomer or multimers determination, (2) radius of gyration (RG) calculation, (3) estimation of pairwise distance distribution, and SAXS quality assessment (e.g., aggregation or not), which provided the basis for further SAXS-assisted model generation and evaluation. And the domain prediction of the target was collected. Then MULTICOM generated the models for the target as follows.

MULTICOM collected all CASP13 server models for each target, and ranked the models using our deep learning-based large-scale quality assessment methods (see details in our CASP13 abstract entitled "Large-scale integration of protein model quality assessment methods using deep learning and contact predictions"). The consensus residue-residue contacts extracted from 50 top ranked models and predicted contacts from DNCON2[1] were combined to generate distance restraints between residues. We generated 1000 new decoys using de novo protein structure prediction[2] with a modified energy function by including new SAXS energy and contact energy terms. The energy function was defined as:

$$E_{total} = E_{force} + E_{contact} + E_{saxs}$$

$$E_{force} = w_{sc-sc} * E_{sc-sc} + w_{sc-pep} * E_{sc-pep} + w_{pep-pep} * E_{pep-pep}$$

$$E_{saxs} = w_{saxsFit} * \frac{\sum_{i=1}^{N} \left| I_{exp}(q_i) - I_{model}(q_i) \right|}{\sum_{i=1}^{N} \left| I_{exp}(q_i) \right|} + w_{saxsKL} \sum_{i=1}^{N} Pr_{model}(r_i) * log \frac{Pr_{model}(r_i)}{Pr_{exp}(r_i)} + w_{saxsRG} * \frac{\left| RG_{exp} - RG_{model} \right|}{\left| RG_{exp} \right|}$$

The SAXS restraints used in the energy function includes: (1) the goodness-of-fitting of a SAXS profile and a model, (2) Kullback-Leibler divergence between the pairwise atom-atom distance (PDF) distributions of a SAXS profile and a model, and (3) the agreement of radius of gyrations of SAXS data and a model.

Moreover, for multi-domain targets (i.e. S0999), we applied the domain-based model evaluation to get the top ranked models for individual domains and performed a SAXS-guided domain assembly to generate full-length protein structural models. The energy function for the domain assembly is defined as:

$$E_{total} = E_{force}^{(intra\ dom)} + E_{force}^{(inter\ dom)} + E_{force}^{(linker)} + E_{saxs}^{(full-length)}$$

During the domain assembly, only the conformation of linker regions was resampled, while the conformation of each domain was kept fixed. So $E_{force}^{(intra\ dom)}$ remained constant during optimization.

Finally, MULTICOM generated 4 lists of scores for all the models, representing SAXS-fitting satisfaction (e.g., χ score of pairwise profile fitting[3]), absolute deviation of RG value, Kullback-Leibler divergence between PDF distributions, and global quality scores. All the four scores were converted to Z-scores. The sum of Z-scores of each model was calculated to rank all the models. The top 5 models were refined by 3Drefine[4] and the local quality scores predicted by ModFOLDclustQ[5] were added into them before they were submitted to CASP13.

1. Adhikari, B., Hou, J. & Cheng, J. DNCON2: Improved protein contact prediction using two-level deep convolutional neural networks. *Bioinformatics* **34**, 1466-1472 (2017).
2. Bhattacharya, D., Cao, R. & Cheng, J. UniCon3D: de novo protein structure prediction using united-residue conformational search via stepwise, probabilistic sampling. *Bioinformatics* **32**, 2791-2799 (2016).
3. Schneidman-Duhovny, D., Hammel, M. & Sali, A. FoXS: a web server for rapid computation and fitting of SAXS profiles. *Nucleic acids research* **38**, W540-W544 (2010).
4. Bhattacharya, D. & Cheng, J. 3Drefine: Consistent protein structure refinement by optimizing hydrogen bonding network and atomic‐level energy minimization. *Proteins: Structure, Function, and Bioinformatics* **81**, 119-131 (2013).
5. McGuffin, L. & Roche, D. Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments. *Bioinformatics* **26** (2010).

CASP13 Tertiary Structure Prediction by the MULTICOM Human Group

Jie Hou[1], Tianqi Wu[1], Renzhi Cao[2], and Jianlin Cheng[1]

1- Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65211, USA 2- Department of Computer Science, Pacific Lutheran University, Tacoma, WA 98447, USA

chengji@missouri.edu

Our MULTICOM human tertiary structure preditor[1] used an automated two-level deep learning scheme to integrate multiple model quality assessment metrics and residue-residue contact predictions to rank and select CASP13 server models as the starting point. Domain-based model evaluation was applied to individual domains of multi-domain targets. The ranking of the top models may be slightly adjusted by human inspection. The model combination approach[2] or model refinement by 3Drefine[3] were applied to top five ranked models. The top models of individual domains of multi-domain targets were combined together by the domain assembly.

Methods

Given a pool of all CASP13 server models for each target, MULTICOM started with filtering out redundant models with high similarity from the same group. Then the models were evaluated by 13 complementary model quality metrics derived from contact predictions by DNCON2[4], single-model quality assessments (i.e. SBROD, OPUS_PSP[5], Model evaluator[6], RF_CB_SRS_OD[7], Rwplus[8], QMEAN[9] and Voronota[10]), and multi-model quality assessments (i.e. Pcons[11], Apollo[12]). These quality scores were used as input for our new deep learning method to generate a consensus ranking of models of each target (for details, see our CASP13 abstract entitled “Large-scale integration of protein model quality assessment using deep learning and contact prediction”). The top five models may be slightly adjusted by human inspection, considering contact predictions, disorder predictions, and pairwise model similarity. Finally, MULTICOM used a model combination approach[2] to combine each of best 5 models with other top ranked similar models together as a potential final model according to the consensus ranking. If the combined model was substantially different from the original model (i.e. GDT-TS < 0.88), 3Drefine[3] method was used to refine the original model to generate a final model instead. If a protein was parsed into multiple domains, the same protocol above was applied to each domain separately, and top 5 models of individual domains were combined into five final full-length models.

Results

We preliminarily evaluated the performance of MULTICOM human predictor along with CASP13 server predictors on 11 CASP13 human targets whose structures were released by the time of writing this abstract. The sum of Z-scores of the first (i.e. TS1) models predicted by these predictors for the 11 targets is reported in **Table 1**. The Z-score of a model was calculated as the model's GDT-TS score minus the average GDT-TS score of all the models in the model pool of a target divided by the standard deviation of GDT-TS scores. A negative Z score was converted to 0 during the summation of Z-scores for a predictor. The results show that MULTICOM performed better than the best server predictors on these targets.

Table 1. The top 10 predictors according to the sum of the Z scores on 11 CASP13 human targets. MULTICOM* was our human predictor, while all others were server predictors. The 11 targets are T0953s1, T0953s2, T0954, T0955, T0958, T0960, T0963, T0965, T0966, T1009 and T1016.

Predictor name	**Sum of Z-scores**	**Predictor name**	**Sum of Z-scores**
MULTICOM*	12.034	BAKER-ROSETTASERVER	8.204
Zhang-Server	10.624	MULTICOM_cluster	7.802
QUARK	10.220	Zhou-SPOT-3D	7.599
RaptorX-DeepModeller	9.662	FALCON	7.578
RaptorX-TBM	9.448	IntFOLD5	7.320

1. Cao, R., Bhattacharya, D., Adhikari, B., Li, J. & Cheng, J. Large-scale model quality assessment for improving protein tertiary structure prediction. Bioinformatics 31, i116-i123 (2015).
2. Wang, Z. MULTICOM: a multi-level combination approach to protein structure prediction and its assessments in CASP8. Bioinformatics 26, doi:10.1093/bioinformatics/btq058 (2010).
3. Bhattacharya, D., Nowotny, J., Cao, R. & Cheng, J. 3Drefine: an interactive web server for efficient protein structure refinement. Nucleic acids research, gkw336 (2016).
4. Adhikari, B., Hou, J. & Cheng, J. DNCON2: Improved protein contact prediction using two-level deep convolutional neural networks. Bioinformatics (2017).
5. Lu, M., Dousis, A. D. & Ma, J. OPUS-PSP: an orientation-dependent statistical all-atom potential derived from side-chain packing. Journal of molecular biology 376, 288-301 (2008).
6. 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 (2009).
7. Rykunov, D. & Fiser, A. Effects of amino acid composition, finite size of proteins, and sparse statistics on distance ‐ dependent statistical pair potentials. Proteins: Structure, Function, and Bioinformatics 67, 559-568 (2007).
8. Zhang, J. & Zhang, Y. A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. PloS one 5, e15386 (2010).
9. Benkert, P., Tosatto, S. C. & Schomburg, D. QMEAN: A comprehensive scoring function for model quality assessment. Proteins: Structure, Function, and Bioinformatics 71, 261-277 (2008).
10. Olechnovič, K. & Venclovas, Č. Voronota: A fast and reliable tool for computing the vertices of the Voronoi diagram of atomic balls. Journal of computational chemistry 35, 672-681 (2014).
11. Wallner, B. & Elofsson, A. Identification of correct regions in protein models using structural, alignment, and consensus information. Protein Science 15, 900-913 (2006).
12. Wang, Z. APOLLO: a quality assessment service for single and multiple protein models. Bioinformatics 27, doi:10.1093/bioinformatics/btr268 (2011).

Large-scale integration of protein model quality assessment using deep learning and contact predictions

Jie Hou and Jianlin Cheng

Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65211, USA

chengji@missouri.edu

Model evaluation plays an important role in protein structure prediction. Different model quality assessment (QA) methods evaluate the quality of protein models from different aspects and have different advantages and limitations. Integrating the power of multiple complementary QA methods has good potentials for improving the accuracy of model quality assessment. Moreover, as the accuracy of contact prediction based on deep learning and co-evolutionary analysis continues to rise, it is useful to include predicted contacts into protein model quality assessment. Thus, we developed two *deep learning* consensus QA methods (MULTICOM_cluster and MULTICOM_construct) to integrate multiple QA methods and contact predictions for predicting the global quality of stage1 and stage2 models of CASP13 targets.

Methods

Our methods used a *deep neural network* to integrate the features generated by 13 QA or contact prediction methods to make quality prediction. Given a pool of models, it first applied the 13 methods whose software were available to generate the input features for each model, which included 10 single-model methods (i.e. DNCON2[1], SBROD, OPUS_PSP[2], RF_CB_SRS_OD[3], Rwplus[4], DeepQA[5], ProQ2[6], ProQ3[7], Dope[8] and Voronota[9]) and three multi-model QA methods (i.e. APOLLO[10], Pcons[11], and ModFOLDclust2[12]). All the 13 methods except DNCON2 are QA methods. DNCON2 is a protein contact predictor, which was used to predict contacts for a target. The percentage of predicted contacts (i.e. short-range, medium-range and long-range contacts) existing in a model of the target was used as a feature. All the input features were used by the deep neural network to predict the quality of each model. The deep neural network was trained on the models of CASP8-11 experiments. 10 trained deep neural networks were obtained from 10-fold cross-validation. All input features of each model were fed into the 10 trained networks to generate 10 quality scores.

MULTICOM_cluster combined the 10 predicted quality scores with the initial input features of 13 QA methods as input for another deep neural network to predict the final quality score, while MULTICOM_construct simply averaged the 10 scores as the final quality score. Prior to the CASP13 experiment, the two methods were benchmarked on the CASP12 dataset and showed a significant improvement compared to the individual QA methods used to generate input features.

Results

We preliminarily evaluated the global quality assessment performance of MULTICOM_cluster and MULTICOM_construct on 14 CASP13 targets whose structures were released by the time of writing this abstract. We used two metrics (i.e. average per target correlation and average per target loss) to assess the quality scores predicted by the two servers against the real quality scores. The loss for each target was calculated as the absolute difference of the GDT-TS score between top 1 model ranked by predicted scores and the overall best model in the model pool. The results are reported in **Table 1**. The results show that MULTICOM_cluster that has a second level of *deep learning* integration worked better than MULTICOM_construct that used the simple averaging.

Table 1. The average per-target correlation score and average loss for global quality assessment of stage1 and stage2 models. The 14 targets are: T0950, T0951, T0953s1, T0953s2, T0954, T0955, T0958, T0960, T0963, T0965, T0966, T0971, T1009 and T1016.

Server name	**Num. of Targets**	**Ave. Corr. Stage1**	**Ave. Corr. Stage2**	**Ave. Loss. Stage1**	**Ave. Loss. Stage2**
MULTICOM_cluster	14	0.828	0.918	0.010	0.036
MULTICOM_construct	14	0.764	0.873	0.010	0.073

1. Adhikari, B., Hou, J. & Cheng, J. DNCON2: Improved protein contact prediction using two-level deep convolutional neural networks. *Bioinformatics* (2017).
2. Lu, M., Dousis, A. D. & Ma, J. OPUS-PSP: an orientation-dependent statistical all-atom potential derived from side-chain packing. *Journal of molecular biology* **376**, 288-301 (2008).
3. Rykunov, D. & Fiser, A. Effects of amino acid composition, finite size of proteins, and sparse statistics on distance - dependent statistical pair potentials. *Proteins: Structure, Function, and Bioinformatics* **67**, 559-568 (2007).
4. Zhang, J. & Zhang, Y. A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. *PloS one* **5**, e15386 (2010).
5. Cao, R., Bhattacharya, D., Hou, J. & Cheng, J. DeepQA: improving the estimation of single protein model quality with deep belief networks. *BMC bioinformatics* **17**, 495 (2016).
6. Uziela, K. & Wallner, B. ProQ2: Estimation of Model Accuracy Implemented in Rosetta. *Bioinformatics*, **btv767** (2016).
7. Uziela, K., Shu, N., Wallner, B. & Elofsson, A. ProQ3: Improved model quality assessments using Rosetta energy terms. *Scientific reports* **6**, 33509 (2016).
8. Shen, M. y. & Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein science* **15**, 2507-2524 (2006).
9. Olechnovič, K. & Venclovas, Č. Voronota: A fast and reliable tool for computing the vertices of the Voronoi diagram of atomic balls. *Journal of computational chemistry* **35**, 672-681 (2014).
10. Wang, Z., Eickholt, J. & Cheng, J. APOLLO: a quality assessment service for single and multiple protein models. *Bioinformatics* **27**, 1715-1716 (2011).
11. Wallner, B. & Elofsson, A. Identification of correct regions in protein models using structural, alignment, and consensus information. *Protein Science* **15**, 900-913 (2006).
12. McGuffin, L. J. & Roche, D. B. Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments. *Bioinformatics* **26**, 182-188 (2009).

MULTICOM_novel (QA)

Deep convolutional neural networks for predicting the quality of single protein structural model

Jie Hou and Jianlin Cheng

Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65211, USA

chengji@missouri.edu

Predicting the global quality and local (residual-specific) quality of a single protein structural model is important for protein structure prediction and application. In CASP13, we benchmarked our new *deep convolutional network* method of predicting the local and global quality of a single model of a protein of arbitrary length. Furthermore, we used a novel multi-task learning framework to study whether global and local quality predictions can synergistically interact to improve prediction performance. Our MULTICOM_novel server based on the method is a single-model quality assessment (QA) method that uses *deep convolutions* to automatically generate input features from a single model to predict its quality, which is different from existing methods relying on hand-crafted features.

Methods

MULITCOM-NOVEL used a novel 1D convolutional neural networks for predicting the local and global quality of a single protein model. Instead of using fixed-size sliding windows to generate features for each residue, the network accepted the input of an entire protein model of arbitrary sequence length and therefore it was able to utilize the global structural information to predict the quality of a specific residue (**Figure 1**).

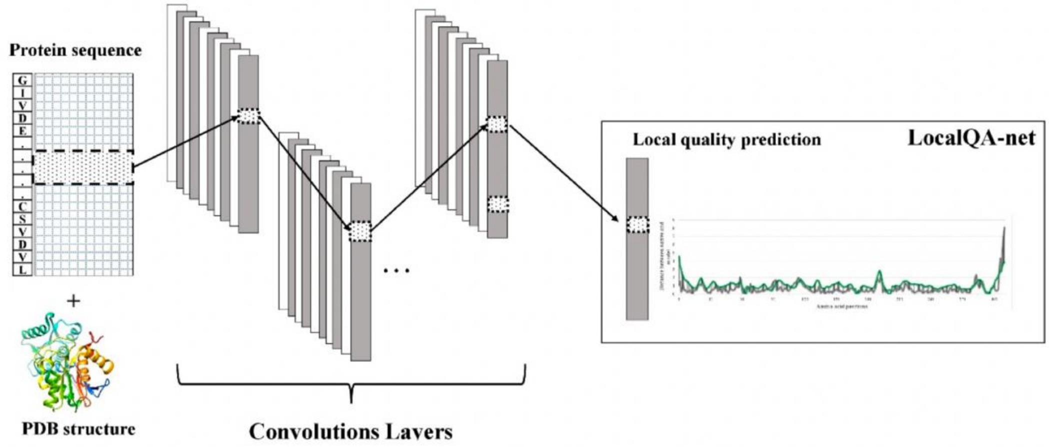


Figure 1. Deep convolutional neural network to predict the local quality of each residue in a model.

The deep network took the following residue-wise raw features and several global features as input, which included (1) amino acid encoding of each residue, (2) position specific scoring matrix (PSSM) profile of each residue derived from the multiple sequence alignment of the protein, (3) predicted secondary structure of each residue, (4) predicted solvent accessibility of each residue, (5) predicted disorder state of each residue, (6) the agreement between the secondary structure of each residue in the model and the predicted one, (7) the agreement of solvent accessibility of each residue in the model and the predicted one, (8) Rosetta energies of each residue as in the ProQ3[1], which was calculated from Van der Waals, side-chains, Hydrogen bonds, and Backbone information, and (9) six global knowledge-based potentials or features of the entire model produced by ModelEvaluator[2], Dope[3], RWplus[4], Qprob[5], GOAP[6], and Surface score .

We designed a training pipeline to integrate local and global quality prediction together, which improved the accuracy of global quality prediction. The method was trained on several large datasets consisting of models from the previous CASP experiments. Overall, the method performed comparably to the state-of-the-art methods in the past CASP11 and CASP12 experiments. The results demonstrate that *1D deep convolutional neural networks* are promising techniques for protein model quality assessment.

Results

We preliminarily evaluated the global quality assessment performance of MULTICOM_novel on 14 CASP13 targets whose structures were released by the time of writing this abstract. We used two metrics (i.e. average per target correlation and average per target loss) to assess the quality scores predicted by our server against the real quality scores. The loss for each target was calculated as the absolute difference of GDT-TS score between top 1 model ranked by predicted scores and the overall best model in the model pool. The results are reported in **Table 1**.

Table 1. The average per-target correlation score and average loss for global quality assessment of stage1 and stage2 models. The 14 targets are: T0950, T0951, T0953s1, T0953s2, T0954, T0955, T0958, T0960, T0963, T0965, T0966, T0971, T1009 and T1016.

Server name	Ave. Corr. Stage1	Ave. Corr. Stage2	Ave. Loss. Stage1	Ave. Loss. Stage2
MULTICOM_novel	0.65	0.61	0.058	0.066

Availability: the source code is available at https://github.com/multicom-toolbox/DeepCovQA.

1. Uziela, K., Shu, N., Wallner, B. & Elofsson, A. ProQ3: Improved model quality assessments using Rosetta energy terms. *Scientific reports* **6**, 33509 (2016).
2. 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).
3. Shen, M. y. & Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein science* **15**, 2507-2524 (2006).
4. Zhang, J. & Zhang, Y. A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. *PloS one* **5**, e15386 (2010).
5. Cao, R. & Cheng, J. Integrated protein function prediction by mining function associations, sequences, and protein–protein and gene–gene interaction networks. *Methods* **93**, 84-91 (2016).
6. Zhou, H. & Skolnick, J. GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. *Biophysical journal* **101**, 2043-2052 (2011).

MULTICOM_cluster, MULTICOM_construct, MULTICOM_novel, DNCON3, MULTICOM (RR)

Deep convolutional neural networks for improving protein contact prediction

Tianqi Wu[1], Jie Hou[1], Badri Adhikari[2], J. Cheng[1]

1- *Department of Electrical Engineering and Computer Science, University of Missouri, Columbia; 2- Department of Mathematics and Computer Science, University of Missouri, St. Louis*

Deep learning and residue-residue co-evolutionary analysis were the two major drivers that advanced protein contact prediction in the last several years. In CASP13, we tested four server predictors and one human predictor based on our deep convolutional neural network method[1] in contact prediction.

Methods

Our four server predictors (MULTICOM_cluster, MULTICOM_construct, MULTICOM_novel, and DNCON3) are based on the same two-level deep convolutional neural networks developed in our DNCON2 method[1] (**Fig. 1**), while they differ in how they generate multiple sequence alignments for co-evolutionary feature generation and how they deal with multi-domain protein targets. **Fig. 1(A)** illustrates a basic convolutional neural network (ConvNet) that converts pairwise residue-residue input information stored in matrices (e.g. residue-residue co-evolutionary features calculated by CCMpred[2], FreeContact[3] and PSICOV[4] and predicted secondary structures) into a predicted residue-residue contact probability matrix at a specific threshold. It has six convolutional layers, each of which has 16 5×5 filters, to transform input L×L matrices into L×L feature maps through convolutions. The feature maps of the 6th convolutional layer are used as input for a filter in the final output convolutional layer to predict a L×L contact probability matrix. **Fig. 1(B)** depicts the two-level convolutional neural networks for contact prediction. At the first level, the input matrices are used by five ConvNets to predict contact maps at five distance thresholds: 6 Å, 7.5 Å, 8 Å, 8.5 Å and 10 Å. At the second level, the five predicted contact probability maps and the original input matrices are used by a ConvNet to predict the final contact map at 8 Angstrom threshold. The whole network was trained on 1426 proteins with known contact maps[1].

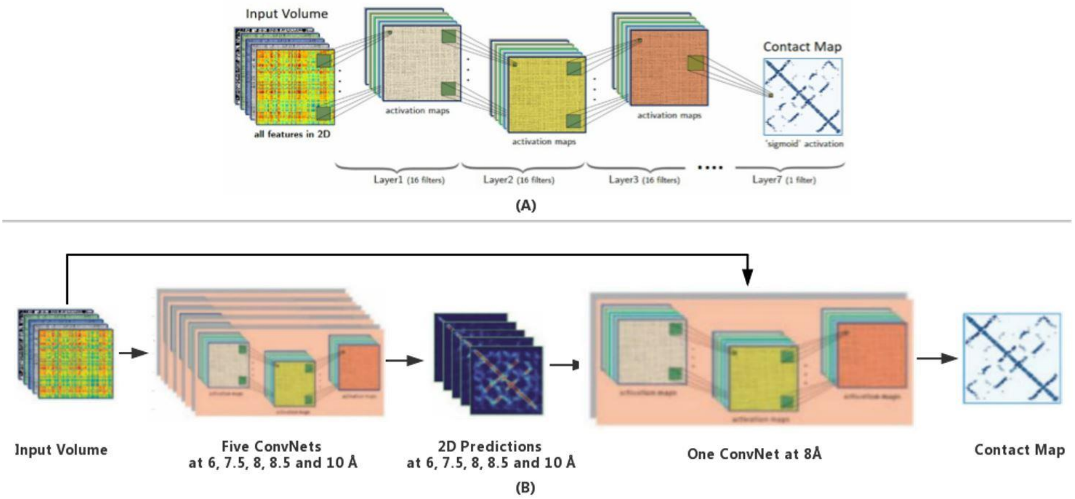


Fig 1. **(A)** A basic convolutional neural network (ConvNet) to predict contact maps at a specific threshold; **(B)** the two-level architecture for predict final contact maps at 8 Angstrom threshold.

In CASP13, we applied four different alignment and domain combination strategies to prepare multiple sequence alignments to generate co-evolutionary input features for the four contact predictors based on the deep learning architecture above. MULTICOM_novel used JackHMMER to search a target against the UniRef database to generate multiple sequence alignments without splitting the target into domains.

MULTICOM_construct and MULTICOM_cluster first predicted domain boundaries for a target. For the full target and each presumably hard domain if exists, the alignments were generated from both HHblits search and JackHMMER search. The two sets of alignments for either the full target or each hard domain are combined. MULTICOM_cluster applied an extra step to remove highly similar (redundant) sequences in the alignment. The alignments were then used to generate co-evolutionary input features for the deep learning networks to predict contact maps for the full target or each hard domain. The predicted contact maps for the full target and each hard domain (if exists) are combined into the final predicted contact map for the target. DNCON3 used the same domain splitting methods as MULTICOM_construct, but it used PSIBLAST and JackHMMER to generate alignment for a full-length target and each hard domain. Our human predictor MULTICOM used the average prediction of the four servers as its prediction.

Results

The performance of our servers was evaluated on the free-modeling (FM) targets of CASP10, 11 and 12 experiments prior to CASP13 experiment. The average precision of the top L/5 long-range (sequence separation ≥24) contact predictions of MULTICOM_cluster, MULTICOM_construct, MULTICOM_novel and the original baseline method - DNCON2 is shown in **Table 1**.

Table 1. The results on CASP10, CASP11 and CASP12 FM datasets

FM Dataset	Domain Count	Precision of top L/5 long-range contact predictions (%)			
		DNCON2	CLUSTER	CONSTRUCT	NOVEL
CASP10	15	36.9	52.4	50.8	48
CASP11	28	51.7	53.8	55.3	53.1
CASP12	28	52.6	56.2	54.3	52.2

Availability

The deep learning code of DNCON2 is available at https://github.com/multicom-toolbox/DNCON2.

1. Adhikari,B. & Cheng,J. (2018). DNCON2: Improved protein contact prediction using two-level deep convolutional neural networks. *Bioinformatics*, **34**, 2018, 1466–1472.
2. Seemayer,S., Gruber,M., & Söding,J. (2014). CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. *Bioinformatics*, **30**, 3128-3130.
3. Kaján,L., Hopf,T. Kalaš,M., Marks,D. & Rost,B. (2014) FreeContact: fast and free software for protein contact prediction from residue co-evolution. *Bioinformatics*, **15**, 85.
4. 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.

Improving Protein Tertiary Structure Prediction by Deep Learning, Contact Prediction and Domain Recognition

Jie Hou, Tianqi Wu, and Jianlin Cheng

Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65211, USA

chengji@missouri.edu

Several interesting technological advances such as deep learning and contact predictions were made to improve template-based or template-free (ab initio) structure modeling in the last few years. In the CASP13 experiment, we improved our protein structure server predictors (MULTICOM_cluster, MULTICOM_construct and MULTICOM_novel) in several aspects: (1) new deep learning-based methods added to improve template identification for hard targets; (2) domain detection by integrating template information and multiple sequence alignments derived from the large sequence database; (3) contact-based ab initio modeling for template-free targets by integrating DNCON2[1] contact predictions with multiple ab initio modeling methods (i.e. CONFOLD[2], FUSION[3], ROSETTA[4], Unicon3d[5]); and (4) the large-scale model quality assessment empowered by deep learning.

Methods

Our three servers (MULTICOM_cluster, MULTICOM_construct and MULTICOM_novel) used a similar protocol to generate a pool of models for a target. It consisted of the following steps: (1) template identification for a target by sequence-sequence alignment, sequence-profile alignment, profile-profile alignment[6] and deep learning-based fold classification; (2) target-templates alignment generation by multiple alignment methods; (3) domain recognition, re-modeling, and assembly, where domains were detected based on both alignments with templates and multiple sequence alignments from the non-redundant sequence database; (4) model generation by template-based modeling; and (5) template-free modelling[4] applied to generate models for targets without reliable templates. For template-based modeling, each of three servers generates about 150-200 models. For free-template modeling, we used contact constraints predicted by DNCON2[1] with four ab initio modeling tools (i.e. CONFOLD[2], FUSION[3], ROSETTA[4], Unicon3d[5]) to generate models separately.

In order to benchmark the influence of contact prediction on tertiary structure modeling, we let MULTICOM_novel run contact-based ab initio structure prediction for up to most 2.5 days on a high-performance computing cluster, whereas MULTICOM_cluster and MULTICOM_construct generally finished the prediction within 2 days on a moderate computer server. For multi-domain targets, the same prediction protocol was applied to each domain, and the top selected conformations of all the domains were combined into full-length models. In total, around 150-250 models were collected for quality assessment.

For model quality assessment, MULTICOM_cluster ranked models primarily based on pairwise similarity between models[7]. MULTICOM_construct and MULTICOM_novel selected best five models based on our two new *deep learning*-based consensus ranking methods (referring to our QA abstract entitled “Large-scale integration of protein model quality assessment using deep learning and contact predictions”) through integration of different quality assessment methods[8] and contact predictions.

Results

We evaluated our three servers on 14 CASP13 targets whose experimental structures were released to date. **Table 1** reports the average GDT-TS scores and TM-scores of top 1 and best of top 5 models.

Table 1. The average GDT-TS scores and TM-scores of top one and best of five models on 14 CASP13 targets. These targets are T0950, T0951, T0953s1, T0953s2, T0954, T0955, T0958, T0960, T0963, T0965, T0966, T0971, T1009 and T1016.

Predictor	Top One		Best of Five	
	GDT-TS	TM-score	GDT-TS	TM-score
MULTICOM_cluster	0.534	0.614	0.546	0.627
MULTICOM_construct	0.518	0.601	0.542	0.623
MULTICOM_novel	0.501	0.576	0.521	0.599

Availability: the source code of several tools of MULTICOM servers is available here: https://github.com/multicom-toolbox .

1. Adhikari, B., Hou, J. & Cheng, J. DNCON2: Improved protein contact prediction using two-level deep convolutional neural networks. *Bioinformatics* **34**, 1466-1472 (2017).
2. Adhikari, B. & Cheng, J. CONFOLD2: improved contact-driven ab initio protein structure modeling. *BMC bioinformatics* **19**, 22 (2018).
3. Bhattacharya, D. & Cheng, J. De novo protein conformational sampling using a probabilistic graphical model. *Scientific reports* **5**, 16332 (2015).
4. Leaver-Fay, A. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol* **487**, doi:10.1016/b978-0-12-381270-4.00019-6 (2011).
5. Bhattacharya, D., Cao, R. & Cheng, J. UniCon3D: de novo protein structure prediction using united-residue conformational search via stepwise, probabilistic sampling. Bioinformatics 32, 2791-2799 (2016).
6. Li, J., Adhikari, B. & Cheng, J. An improved integration of template-based and template-free protein structure modeling methods and its assessment in CASP11. *Protein and peptide letters* **22**, 586-593 (2015).
7. Wang, Z., Eickholt, J. & Cheng, J. APOLLO: a quality assessment service for single and multiple protein models. *Bioinformatics* **27**, 1715-1716 (2011).
8. Cao, R., Bhattacharya, D., Adhikari, B., Li, J. & Cheng, J. Massive integration of diverse protein quality assessment methods to improve template based modeling in CASP11. *Proteins: Structure, Function, and Bioinformatics* **84**, 247-259 (2016).

PconsC4: fast, free, easy, and accurate contact predictions

D. Menéndez Hurtado[1], M. Michel[1], and A. Elofsson[1]

1 - Science for Life Laboratory and Department of Biochemistry and Biophysics, Stockholm University

arne@bioinfo.se

PconsC4[1] is a contact predictor designed to be fast, as it only requires one alignment as input and is easy to install, thanks to the lack of external requirements; while yielding state of the art results.

It combines the statistical power of Direct Coupling Analysis (DCA) for finding the causes for the observed correlated mutations in large multiple sequence alignments; with the pattern recognition abilities of a deep network, capable of extracting the underlying signal even for shorter families.

Methods

PconsC4 combines an accurate global statistical model, GaussDCA[2], with more sensitive but noisier local statistics, including mutual information, and cross entropy. All the features are fed to a deep convolutional network based on the U-net[3] architecture to extract patterns and clean the predictions. The network is trained to predict the probability of contacts at 6, 8, and 10 Å thresholds, as well as the distance between residues; of which only the 8 Å contacts were submitted to CASP.

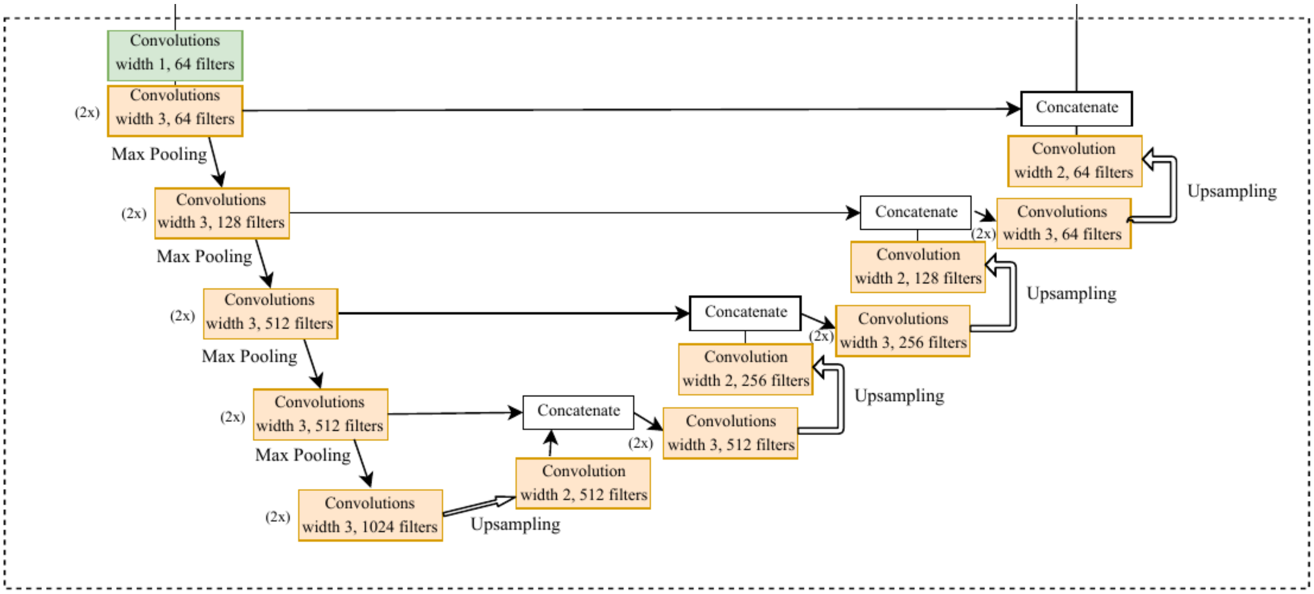

Convolutions width 1, 64 filters
(2x)
Convolutions width 3, 64 filters
Max Pooling
(2x)
Convolutions width 3, 128 filters
Max Pooling
(2x)
Convolutions width 3, 512 filters
Max Pooling
(2x)
Convolutions width 3, 512 filters
Max Pooling
(2x)
Convolutions width 3, 1024 filters
Upsampling
Convolution width 2, 512 filters
Concatenate
(2x)
Convolutions width 3, 512 filters
Upsampling
Convolution width 2, 256 filters
Concatenate
(2x)
Convolutions width 3, 256 filters
Upsampling
Convolution width 2, 128 filters
Concatenate
(2x)
Convolutions width 3, 64 filters
Upsampling
Convolution width 2, 64 filters
Concatenate

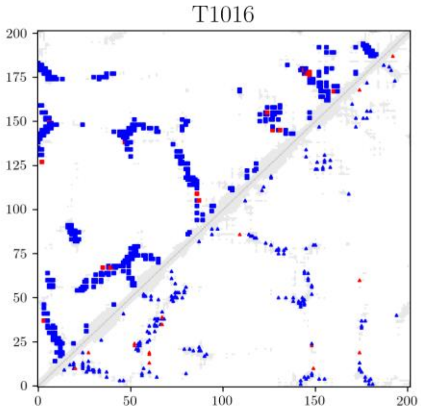
Figure 1. Schematic representation of the U-net architecture. All inputs are concatenated and passed through the upper left corner.

As a comparison, we provide as well the contact probabilities provided by GaussDCA. The raw scores provided by the method were transformed into probabilities by fitting a simple sigmoid function to a small sample. It should be noted that this is a purely statistical method based on the alignment columns, unaware of the separation between residues. This is why we sometimes get low contact probabilities for residues next to each other in the sequence.

The alignments were generated by five iterations of jackhmmer[4], with a E-value threshold of 1 on Uniref 90. No further effort was done to improve this stage.

Results

We present a comparison between GaussDCA and PconsC4 on the target T1016. The alignment is composed of 73151 hits. GaussDCA (bottom half) has a Top-L PPV of 0.92. With PconsC4 (upper half) we can increase the coverage up to 2.5 L, while also improving the PPV to 0.96.



Availability

PconsC4 is freely available under the GPL license from https://github.com/ElofssonLab/PconsC4. Installation is easy using the pip command and works on any system with Python 3.5 or later and a modern GCC compiler.

1. Michel M., Menendez Hurtado D., Elofsson A. (2018), PconsC4: fast, free, easy, and accurate contact predictions. *bioRxiv*
2. Baldassi, C., Zamparo, M., Feinauer, C., Procaccini, A., Zecchina, R., Weigt, M., and Pagnani, A. (2014). Fast and accurate multivariate gaussian modeling of protein families: Predicting residue contacts and protein-interaction partners. *PLOS ONE*, **9** (3), 1–12.
3. Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *MICCAI 2015*, pages 234–241
4. Johnson L. S., Eddy S. R., and Portugaly E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* **11**, 431

Protein Structure Modelling Software with Machine Learning-Based Scoring Functions

Toshiyuki Oda

odat1248@gmail.com

PepBuilderJ is a newly developed modelling software which can do comparative modelling and threading.

Methods

PepBuilderJ uses sampled sidechain rotamers and backbone conformers to build peptide chains and machine learning-based scoring functions to find native-like structures. For CASP13, I implemented small decision trees considering their computation speed. They produce residue-level scores using distance information between atoms in modelled chains as input features. The scoring functions are used for loop modelling, threading, and prioritizing the models for submission. As several bugs and insufficient points were found and fixed during the season, the following procedures didn't work well for all (especially early) targets. The comprehensive performance and detailed explanation about this software and the following protocols will be done in later on.

For the comparative modelling, I used hhsuite[1] to search for templates and make alignments. The HMM database for structural domains defined by the PDP[2] algorithm in chains from PDB(http://www.rcsb.org/)[3] were constructed for templates. The structural domains were filtered using cd-hit[4] with 60% identity threshold. To build the HMM profiles, I employed the supervised profile construction method. Because the structures of proteins deposited in PDB are already known, we can know good profiles for hhblits which can find highest number of true positive (TP) hits. TM-score calculated by TM-align[5] 0.5 was used as the threshold of TP and false positive (FP) hits. The hhblits searches were performed iteratively until the number of TP hits, whose e-values were lower than any FP hits, converged. Jackhmmer[6] (HMMER3 version 3.1b2) was used to filter the sequences in the results. The query HMMs were built with 5 iterations of hhblits search against uniprot20 database. The threading was done using the 40% ID filtered subset of SCOPe7 domains (version 2.07) as templates. GROMACS[8] minimization (following the procedure in http://www.bevanlab.biochem.vt.edu/Pages/Personal/justin/gmx-tutorials/lysozyme/ with small modifications) was performed to do the atom-level refinement. However, in situations with limited time resources for GROMACS to run, the step was omitted. The models for submission were selected with manual intervention based on the scores and visualization result by PyMOL[9].

For the multimer targets, all entries in the PDB biounit section (ftp://ftp.wwpdb.org/pub/pdb/data/biounit/PDB/divided/) were downloaded. The entries which contain homologs of the monomer templates (found by hhblits as written above) were extracted as multimer templates using blastp[10]. The e-value threshold was set to 0.0001 or much lower if were there many homologs. The monomer models constructed as written above were aligned using TM-align with the chains in the multimer templates. I discarded alignments whose TM-scores were less than or equal to 0.5 (or 0.3 if none of the multimer templates had remained). After the structural alignments, the number of interactions (distance between two Cβ (or pseudo Cβ for Glycine) less than 6.0Å) and crashes (distance between two Cα was less than 3.5Å) between chains were recorded. The models for submission were selected with manual intervention based on the number of aligned units, the number of interactions, the number of crashes, the scores by PepBuilderJ, and variations of templates. Before the submissions, interface remodeling using the scoring function and GROMACS minimization was performed to remove crashes if there were enough time and computation resources.

*T.O. is an employee of Lifematics Inc. This work was done privately by the author.

1. Remmert, M., Biegert, A., Hauser, A. & Soding, J. (2011). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods*. **9**, 173-175
2. Alexandrov, N. & Shindyalov, I. (2003). PDP: protein domain parser. *Bioinformatics*. **19**, 429-430
3. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. & Bourne, P.E. (2000). The Protein Data Bank. *Nucleic Acids Res*. **28**, 235-242
4. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*. **28**, 3150-3152
5. Zhang, Y. & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res*. **33**, 2302-2309
6. Potter, S.C., Luciani, A., Eddy, S.R., Park, Y., Lopez, R. & Finn, R.D. (2018). HMMER web server: 2018 update. *Nucleic Acids Res*. **46**, W200-W204
7. Fox, N.K., Brenner, S.E. & Chandonia, J.M. (2014). SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res. 42, D304-309
8. Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A.E. & Berendsen, H.J. (2005). GROMACS: fast, flexible, and free. *J Comput Chem*. **26**, 1701-1718
9. Schrodinger, LLC. (2015). The PyMOL Molecular Graphics System, Version 1.8
10. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T.L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics.* **10**, 421

Ab-initio Protein Structure Prediction using MD Simulations

A. Pérez[1] and G. De Fabritiis[1,2]

1 – Computational Science, GRIB-IMIM. Universitat Pompeu Fabra, Barcelona, Spain, 2 – Institució Catalana de Recerca d'Estudis Avançats (ICREA), Barcelona, Spain

gianni.defabritiis@upf.edu

Even though native conformations of proteins are usually represented by static crystal structures, the folding process is a dynamical process. For ab-initio structure prediction, where only the sequence is known, we have focused on protein folding dynamics to perform our predictions.

Molecular dynamics (MD) simulations is the most used method to study protein dynamics, due to their capacity to describe dynamical processes with atomic resolution. Progress in the field of MD during the past decade has reduced the computational cost of simulations, increasing the capacity to reach slow timescales, up to the order of milliseconds[1]. The decreasing computational cost and the advent of Markov State Model analysis[2] has transformed MD simulations into high-throughput experiments, where thousands of short simulations run in parallel. Simultaneously, the development of novel adaptive sampling schemes for high-throughput MD simulations has increased their efficiency, reducing the amount of simulations needed to obtain converged statistics[3-6].

All of the aforementioned developments has brought MD simulations up to a point where fast folding timescales can be reached with unbiased simulations[7]. The unbiased simulation of several folding events has been possible for a short amount of proteins, mostly fast folders. For larger ones, it is mostly impossible due to folding timescales for most proteins and the prediction yielding time being too slow. Nevertheless, recent improvements in contact predictions and adaptive sampling pushed us to test the current capabilities of unbiased MD simulations using contacts and secondary structure as prior information to perform protein folding in this CASP challenge.

Methods

The recent advances in machine learning of contact predictions using evolutionary methods have given rise to novel prediction methods that have demonstrated their predictive power in previous CASP events. Contact prediction is now powerful enough to guide the sampling algorithm to the native state. We have used the predictions provided by RaptorX Contact Prediction[8], a deep learning based contact map prediction tool trained on evolutionary coupling and sequence conservation information. The secondary structure prediction was performed with the PSIPRED server[9].

For CASP we planned to improve our adaptive sampling algorithms using a solid reinforcement learning framework. We used an off-policy version of the exploration-exploitation trade-off described in the multi-armed bandit problem. In the problem, a gambler must choose which arm to play, from a pool of K arms in a slot machine. In multi-armed bandit problems, each arm (simulation) has a different payoff distribution (usefulness), and the gambler (sampler) has to balance exploration to learn which are the most rewarding arms (simulations), and exploitation of the best arms. The goal is to reduce the regret, which is described as the difference between the gambler's total reward and the best arm's total reward over *n* trials[10]. The algorithm rewards sampling from metastable states, but also promotes exploration based on the idea of "optimism in face of uncertainty", pushing sampling to unexplored states. We had only a preliminary version of the method for the time of submission in CASP, but we are now continuing to improve it.

We have performed high-throughput MD simulations for a small subset of 9 CASP targets, containing only proteins with less than 100 residues, for a total amount of 10-50 μs for each target. The simulations were performed with ACEMD[11] using CHARMM22* forcefield[12]. Trajectory analysis and MSM construction were performed with HTMD[5]. The MSM was constructed featurizing coordinates into residue contacts, using tICA[13] for dimensionality reduction and Kmeans clustering to discretize the tICA space. The predicted protein structures were selected, by visual inspection, from conformations coming from the most stable states in a coarse-grained MSM and from top scoring conformations from the goal function.

1. Lindorff-Larsen,K., Maragakis,P., Piana,S., Shaw,D.E. (2016) Picosecond to Millisecond Structural Dynamics in Human Ubiquitin. *J. Phys. Chem. B.* ***120(33)***, 8313-8320.
2. Prinz,J.H., Wu,H., Sarich,M., Keller,B., Senne,M., Held,M., Chodera,J.D., Schütte,M., Noé,F. (2011). Markov models of molecular kinetics: Generation and validation. *J. Chem. Phys.* **134**, 174105/1 – 174105/23.
3. Sinhal,N., Pande,V.S. (2005) Error analysis and efficient sampling in Markovian state models for molecular dynamics. *J. Chem. Phys.* **123,** 204909/1 – 204909/13.
4. Doerr,S. De Fabritiis,G. (2014) On-the-fly Learning and Sampling of Ligand Binding by High-Throughput Molecular Simulations. *J. Chem. Theory Comput.* **10(5)**, 2064-2069
5. Doerr,S., Harvey,M.J., Noé,F., De Fabritiis,G. (2016) HTMD: High-Throughput Molecular Dynamics for Molecular Discovery. *J. Chem. Theory Comput.* **10(5),** 2064-2069.
6. Zimmerman,M.I., Bowman,G.R. (2015) FAST Conformational Searches by Balancing Exploration/Exploitation Trade-Offs. *J. Chem. Theory Comput.* **11(12)** 5747-5757.
7. Lindorff-Larsen,K., Piana,S., Dror,R.O., Shaw,D.E. (2011) How Fast-Folding Proteins Fold. *Science* **334**, 517-520.
8. Wang,S., Sun,S., Li,Z., Zhang,R., Xu,J. (2017) Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLoS Computational Biology* **13(1)**.
9. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292** 195-202.
10. Sutton,R.S., Barto,A.G. (2018) Reinforcement Learning: An Introduction. *MIT Press*, *second edition*.
11. Harvey,M.J., Giupponi,G., De Fabritiis,G. (2009) ACEMD: Accelerating Biomolecular Dynamics in the Microsecond Time Scale. *J. Chem. Theory Comput.* ***5(6)*** 1632-1639.
12. Piana,S., Lindorff-Larsen,K., Shaw,D.E. (2011) How robust are protein folding simulations with respect to force field parameterization? *Biophys. J.* **100** L47-L49.
13. Pérez-Hernández ,G., Paul,F., Giorgino,T., De Fabritiis,G., Noé,F. (2013) Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.* **139(1)** 015102/1-015102/13.

AngularQA: Protein Model Quality Assessment with LSTM Networks

Matthew Conover[1], Max Staples[1], Dong Si[2], and Renzhi Cao[1*]

1 - Department of Computer Science, Pacific Lutheran University, Tacoma, WA 98447, 2 - Division of Computing and Software Systems, University of Washington-Bothell, Bothell, WA 98011

*caora@plu.edu

In CASP 13, we tested our new developed method AngularQA (PLU-AngularQA server) for protein model quality assessment (QA) category. QA plays an important role in protein structure prediction [1]. Traditional protein QA methods suffer from searching databases or comparing with other models for making predictions, which usually fail. We propose a novel protein single-model QA method which is built on a new representation that converts raw atom information into a series of carbon-alpha (Cα) atoms with side-chain information, defined by their dihedral angles and bond lengths to the prior residue. An LSTM network is used to predict the quality by treating each amino acid as a time-step and consider the final value returned by the LSTM cells. To the best of our knowledge, this is the first time anyone has attempted to use an LSTM model on the QA problem; furthermore, we use a new representation which has not been studied for QA [2]. In addition to angles, we make use of sequence properties like secondary structure at each time-step, without using any database.

Methods

For the initial data preparation part, all data used in training our LSTM network comes from 3DRobot decoys [3] and CASP 9, 10, and 11 [4]. These have 92,535, 36,083, 15,901, and 14,193 models respectively from which we draw for training. Validation occurs on the CASP12, of which we use 6,790 models across 40 targets [4]. We begin by filtering all the models. During this process we verify the residue sequences in the predicted structures line up correctly with the native structure, and throw out any predicted models with gaps in the center. In addition, We throw out any models for which we do not have the native structure. After filtering, we are left with a total of 128,439 models with 121,875 training models and 6564 validation models.

After that, we calculate the angles and bond lengths along the backbone and side-chain as was described by UniCon3D[2]. The result is a sequence of angle and bond length information provided for each residue following along the carbon backbone. In addition, we also calculate the proximity counts, which are also calculated by counting the number of Cα atoms within a set radius of each residue's Cα atom. We perform this calculation for all radii in the discrete range [5Å, 15Å]. Moreover, the second structure is parsed by DSSP program[5], but there is no secondary structure prediction used in our method, which is different from a lot of traditional QA methods[6–12]. The machine learning technique is applied to train a LSTM network on the processed feature vectors, and each LSTM cells uses a hyperbolic tangent activation with a hard sigmoid recurrent activation.

Availability

The AngularQA software is available in Github at the following link:
https://github.com/caorenzhi/AngularQA

1. Cao, R., Bhattacharya, D., Adhikari, B., Li, J. & Cheng, J. (2015). Large-scale model quality assessment for improving protein tertiary structure prediction. *Bioinformatics* **31**, i116–23
2. Bhattacharya, D., Cao, R. & Cheng, J. (2016). UniCon3D: de novo protein structure prediction using united-residue conformational search via stepwise, probabilistic sampling. *Bioinformatics* **32**, 2791–2799

3. Deng, H., Jia, Y. & Zhang, Y. (2016). 3DRobot: automated generation of diverse and well-packed protein structure decoys. *Bioinformatics* **32**, 378–387
4. Moult, J., Pedersen, J.T., Judson, R. & Fidelis, K. (1995). A large-scale experiment to assess protein structure prediction methods. *Proteins* **23**, ii–v
5. Joosten, R.P., te Beek, T.A.H., Krieger, E., Hekkelman, M.L., Hooft, R.W.W., Schneider, R., Sander, C. & Vriend, G. (2011). A series of PDB related databases for everyday needs. *Nucleic Acids Res.* **39**, D411–9
6. Uziela, K., Menéndez Hurtado, D., Shu, N., Wallner, B. & Elofsson, A. (2017). ProQ3D: improved model quality assessments using deep learning. *Bioinformatics* **33**, 1578–1580
7. Cao, R. & Cheng, J. (2016). Protein single-model quality assessment by feature-based probability density functions. *Sci. Rep.* **6**, 23990
8. Cao, R., Bhattacharya, D., Hou, J. & Cheng, J. (2016). DeepQA: improving the estimation of single protein model quality with deep belief networks. *BMC Bioinformatics* **17**, 495
9. Manavalan, B. & Lee, J. (2017). SVMQA: support-vector-machine-based protein single-model quality assessment. *Bioinformatics* **33**, 2496–2503
10. Olechnovič, K. & Venclovas, Č. (2017). VoroMQA: Assessment of protein structure quality using interatomic contact areas. *Proteins* **85**, 1131–1145
11. Derevyanko, G., Grudinin, S., Bengio, Y. & Lamoureux, G. (2018). Deep convolutional networks for quality assessment of protein folds. *Bioinformatics* doi:10.1093/bioinformatics/bty494
12. Kryshtafovych, A., Barbato, A., Monastyrskyy, B., Fidelis, K., Schwede, T. & Tramontano, A. (2016). Methods of model accuracy estimation can help selecting the best models from decoy sets: Assessment of model accuracy estimations in CASP11. *Proteins* **84 Suppl 1**, 349–369

TopQA: A Topological Representation for Single-Model Protein Quality Assessment with Machine Learning Technique

John Smith[1], Natalie Stephenson[1], Dong Si[2], and Renzhi Cao[1*]

1 - Department of Computer Science, Pacific Lutheran University, Tacoma, WA 98447, 2 - Division of Computing and Software Systems, University of Washington-Bothell, Bothell, WA 98011

*caora@plu.edu

We tested our recently developed method TopQA (attended CASP 13 as PLU-TopQA server) for protein model quality assessment (QA), which is one of the most important steps for protein structure prediction problem [1]. To the best of our knowledge, TopQA is the first method to tackle protein model quality assessment problem by analyzing the topology of the predicted protein structure. All Carbon Alpha atoms from predicted structure are processed by our method, and the topology of the structure is normalized into a cube representation. With the help of latest machine learning techniques - convolutional neural network (CNN), GDT_TS score is predicted for any given protein structure model. Our TopQA method is single-model QA method, which could be used to produce model quality assessment for any single protein structure model.

Methods

First of all, we prepared the training datasets for developing TopQA. In summary, we used a total of 176 target proteins from the CASP10 and CASP11 datasets (These can be found at: http://predictioncenter.org/download_area/), including 15,901 CASP10 models and 14,139 CASP11 models. Each protein structure model is in PDB format, and provides a standard representation for macromolecular structure data. Traditional methods [2–8] usually use the 3D structure of protein structure model in PDB format directly with help of other properties of protein sequence, but no method has tried to modify the representation of the 3D structure model. We proposed a new representation of the 3D structure model and use that for training machine learning model.

Second, we created our new representation for each of PDB file. The 3D coordinates of each carbon alpha atoms were extracted, and the whole topology of this structure was kept while we scale the structure into a cube with size 1. In addition, this representation systematically mapped the mass of each carbon alpha atom in the backbone of the protein model to a three-dimensional space in the cube. This 1x1x1 cube can be scaled to any size, although for our model we generally used a 52x52x52 (see the results section for more information regarding varying dimensions). Finally, rotations were applied to this new representation to generate model robust model. With this approach, we were able to map each model numerous times, viewing the model from a slightly different angle each time. Normally, it's very costly to apply rotation to the model, but one rotation of each model in our model representation would take a second and would be used in our final representation. Once we formatted the PDB files into this representation, we were left with a 3-dimensional matrix in which every value represented the mass of a single atom in the protein's backbone (several of these values were zero, as the matrix included the empty space of the cube surrounding the protein structure as well as the empty space encapsulated by the structure)

Finally, after transforming the pdb files into our new topologically-based representation, we trained a convolutional neural network (CNN) model. This CNN was made of two convolutional layers, a single pooling layer as well as two dense layers. The CNN was an appealing choice of machine learning method as it lends itself to images and matrices quite well. We have also considered other types of machine learning methods such as an SVM, however, the CNN performed the best.

Availability

The TopQA software is available in Github at the following link:
https://github.com/caorenzhi/TopQA

1. Bhattacharya, D., Cao, R. & Cheng, J. (2016). UniCon3D: de novo protein structure prediction using united-residue conformational search via stepwise, probabilistic sampling. *Bioinformatics* **32**, 2791–2799
2. Uziela, K., Menéndez Hurtado, D., Shu, N., Wallner, B. & Elofsson, A. (2017). ProQ3D: improved model quality assessments using deep learning. *Bioinformatics* **33**, 1578–1580
3. Cao, R. & Cheng, J. (2016). Protein single-model quality assessment by feature-based probability density functions. *Sci. Rep.* **6**, 23990
4. Cao, R., Bhattacharya, D., Hou, J. & Cheng, J. (2016). DeepQA: improving the estimation of single protein model quality with deep belief networks. *BMC Bioinformatics* **17**, 495
5. Manavalan, B. & Lee, J. (2017). SVMQA: support-vector-machine-based protein single-model quality assessment. *Bioinformatics* **33**, 2496–2503
6. Olechnovič, K. & Venclovas, Č. (2017). VoroMQA: Assessment of protein structure quality using interatomic contact areas. *Proteins* **85**, 1131–1145
7. Derevyanko, G., Grudinin, S., Bengio, Y. & Lamoureux, G. (2018). Deep convolutional networks for quality assessment of protein folds. *Bioinformatics* doi:10.1093/bioinformatics/bty494
8. Kryshtafovych, A., Barbato, A., Monastyrskyy, B., Fidelis, K., Schwede, T. & Tramontano, A. (2016). Methods of model accuracy estimation can help selecting the best models from decoy sets: Assessment of model accuracy estimations in CASP11. *Proteins* **84 Suppl 1**, 349–369

Ab initio Protein Folding Using guided by Contact Prediction using Multi-Perspective Deep Convolutional Neural Networks

Badri Adhikari[1], Nitesh Kafle[2], Renzhi Cao[3], Anthony Ackah-Nyanzu[1]

1 - University of Missouri-St. Louis, 2 – Lord Buddha Education Foundation, 3 – Pacific Lutheran University

adhikarib@umsl.edu

We participated in the tertiary structure prediction category (server only) with a contact guided folding pipeline developed based on a new method for contact prediction, and the existing method CONFOLD2[1] for three-dimensional modeling.

Methods

With the open-source tool DNCON2[2] as a reference, we developed four different deep convolutional neural networks (CNNs): 1) a basic deep CNN, 2) a dilated CNN, 3) a separable CNN, and 4) a basic deep CNN trained to predict contacts at the distance thresholds of 6, 8 and 10 Angstroms. These networks were trained and tested on the standard dataset of 1426 proteins discussed in the DNCON2 method. The architecture of the basic deep CNN method consists of 17 layers of 64 filters of size 3x3 and the last output layer with one 3x3 filter. After each layer, the activations are padded with zeros so that the input dimensions (300x300) are maintained through all the layers including the output of the last layer. The basic deep CNN model has a total of 625,921 trainable parameters: 64 3x3 filters on the 56 channels give along with 64 bias values result to 32,320 parameters, 16 layers of 64 3x3 filters on 64 channel activations with 64 bias values at each layer result in a total of 590,848 parameters, 128 parameters for batch normalization at each of the 17 layers result in 2,176 parameters, and one 3x3 filter in the last layer on 64 activation channels along with a bias results in 577 parameters.

The architecture of the dilated CNN consists of 13 regular convolutional layers each with 64 3x3 filters followed by two dilated CNN layers. Both dilated CNN layers consist of 64 3x3 filters with a dilation rate of 2. The last layer is a single 3x3 filter. The dilated CNN model has 551,809 parameters total. Similarly, the separable CNN model has its first layer of 64 3x3 filters, followed by 15 depthwise separable CNN layers (SeparableConv2D in Keras) each with 64 3x3 filters, and the last layer with a depthwise separable CNN with 1 3x3 filter. This model has a total of 101,185 parameters. The fourth model is an extension of the basic deep CNN model trained to predict contacts at the thresholds of 6 and 10 Angstroms at the same time. To achieve this, we replace the last layer with three parallel CNN blocks each consisting of two convolutional layers - first layer with 32 3x3 filters and second with one 3x3 filter as the output. Each of these three outputs separately predict contacts at 6, 8, and 10 Angstroms. When calculating the binary cross entropy loss, we weight these outputs such that the weights are 0.25, 1.0, and 0.25 for 6 Angstroms predictor, 8 Angstroms predictor, 10 Angstroms predictor respectively. Predictions by the four methods are averaged to predict the final set of contacts. Finally, we used top 2L long-range and medium-range contacts (L is the length of a protein) to predict five models using the CONFOLD method.

Results

For contact prediction, when trained using the subset of 1230 proteins and tested on the remaining 196 proteins, the precision of top L/5 long-range contacts ranges from 72.8% to 73% for the four methods. Averaging the predictions of the four methods, we obtain average precision of 75.8% on the 196 proteins, suggesting that the perspective from multiple methods is significantly better than any of the individual methods. We also evaluated our overall method against the experimental structures of some of the targets released so far. While the official CASP results have not been published yet, our preliminary evaluations at target level suggest that the TM-score values of the best-of-five models for the targets T0955, T0958, T0963, T0965, T0971, T1009, and T1016 are 0.36, 0.29, 0.05, 0.17, 0.60, 0.71, and 0.68 respectively.

Availability

The original DNCON2 method and the CONFOLD2 method are publicly available at https://github.com/multicom-toolbox/DNCON2/ and https://github.com/multicom-toolbox/confold2 respectively.

1. Adhikari B, Cheng J. CONFOLD2: improved contact-driven ab initio protein structure modeling. *BMC Bioinformatics*. 2018;19(1):22.
2. Adhikari B, Hou J, Cheng J. DNCON2: improved protein contact prediction using two-level deep convolutional neural networks. *Bioinformatics*. December 2017.

Model Quality Assessment through Deep Learning

D. Menéndez Hurtado, K. Uziela, and A. Elofsson

Science for Life Laboratory and Department of Biochemistry and Biophysics, Stockholm University

arne@bioinfo.se

Deep learning has been a revolution in machine learning since it opened the doors to leverage bigger datasets by allowing us to take advantage of the structure of the data. In this edition, we applied two quality assessment methods: ProQ3D[1], a classic machine learning method that combines a large number of features and a simple feed-forward neural network trained on different target functions[2]; and ProQ4[3], a novel method using a minimal set of inputs and designed to boost per target correlations.

Methods

ProQ3D uses a simple three-layered perceptron that takes as inputs a collection of structural features, such as observed contacts between residues and atoms, sequence-based features, like predicted secondary structure and profiles, and physico-chemical properties, like energy functions computed by Rosetta. We present four versions, trained on different target functions: LDDT, CAD, TM-score, and S-score.

ProQ4 is trained on the same data, but uses a much simpler description: the structural features given by DSSP, dihedral angles, relative surface area, and secondary structure; and simple statistics, such as entropy of each column, from a multiple sequence alignment. The basic architecture is a deep convolutional network.

The main difference between ProQ4 and other methods is that the network is trained in a comparative fashion: at every iteration, two models from the same target were shown, and the network was trained to predict not only the scores of each model, but also which one was better. This is a way of augment our data, and to take advantage of the structure of the problem. A schematic of the network as it was trained is shown on the Figure 1, but for predictions only one of the two symmetrical towers is used (from alignment and model features – 0 to LDDT). This training protocol emphasizes ranking of models inside each target, training in effect on per target correlations.

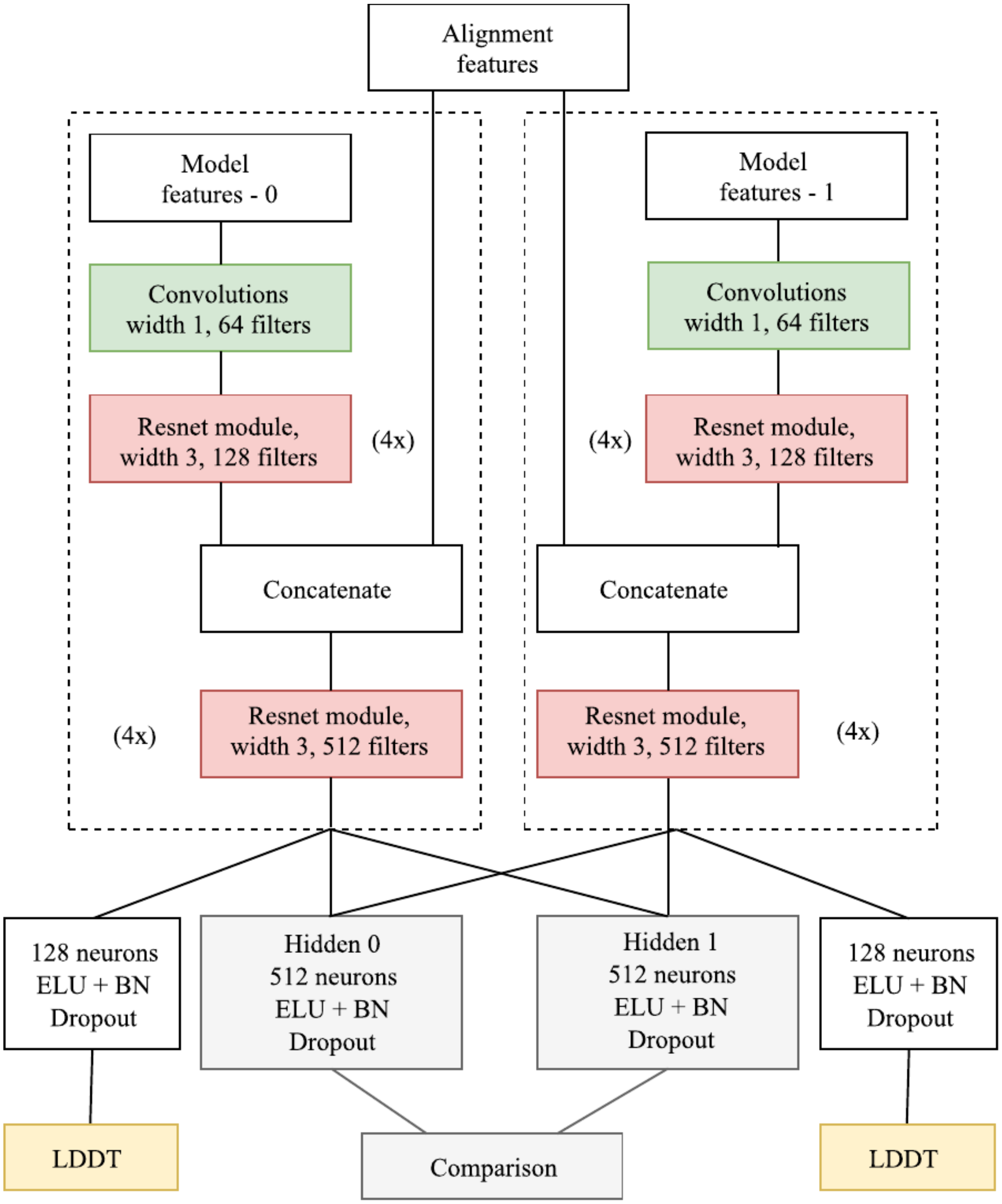

Alignment features
Model features - 0
Convolutions width 1, 64 filters
Resnet module, width 3, 128 filters
(4x)
Concatenate
(4x)
Resnet module, width 3, 512 filters
Model features - 1
Convolutions width 1, 64 filters
(4x)
Resnet module, width 3, 128 filters
Concatenate
Resnet module, width 3, 512 filters
(4x)
128 neurons ELU + BN Dropout
Hidden 0 512 neurons ELU + BN Dropout
Hidden 1 512 neurons ELU + BN Dropout
128 neurons ELU + BN Dropout
LDDT
Comparison
LDDT

Figure 1. Schematic representation of the ProQ4 network as it was trained, to encourage comparisons between models. This network design is optimised for per target correlations. When predicting, only one half is used: alignment features, one of the models, and the corresponding LDDT

Availability

ProQ3D is available as a web server and standalone at proq3.bioinfo.se. ProQ4 can be downloaded from github.com/ElofssonLab/ProQ4.

1. Uziela K., Menéndez Hurtado D., Shu N., Wallner B., Elofsson A. ProQ3D: improved model quality assessments using deep learning, *Bioinformatics*, Volume 33, Issue 10, 15 May 2017, Pages 1578–1580
2. Uziela K., Menéndez Hurtado D., Shu N., Wallner B., Elofsson A. Improved protein model quality assessments by changing the target function (2018). *Proteins*.
3. Menéndez Hurtado D., Uziela K., Elofsson A. Deep transfer learning in the assessment of the quality of protein models (2018), *ArXiv.*

C-QUARK: Ab Initio protein structure folding simulation guided by deep-learning based contact predictions

S M Golam Mortuza[1], Chengxin Zhang[1], Yang Li[1,2], Wei Zheng[1], Yang Zhang[1]

1-Department of Computational Medicine and Bioinformatics, Department of Biological Chemistry, University of Michigan, 100 Washtenaw Ave, Ann Arbor, MI 48109;2-School of Computer Science and Engineering, Nanjing University of Science and Technology, Xiaolingwei 200, Nanjing, China, 210094

yangzhanglab@umich.edu

C-QUARK is new pipeline built on QUARK,[1, 2] which was developed for *ab initio* protein structure prediction by the assembly of continuously sized structural fragments. In C-QUARK (Mortuza et al, in preparation), sequence-based contact-map predictions have been used for constraining the QUARK based structure assembly simulations.

Starting from the query sequence, a set of structural fragments with 1-20 residues is collected from the structure of unrelated proteins in the PDB. Full-length structure models are then assembled from the fragments by replica-exchanged Monte Carlo (REMC) simulations. The original knowledge-based QUARK force field contains a variety of local structure features derived from sequence (e.g. beta-turns, backbone torsion angles, solvent accessibility, and helix and strand packing possibilities). In particular, a set of long-range residue-residue contacts derived from the fragment-based distance profiles were used to assist the fragment assembly simulation.[2] The final models were selected based on the SPICKER clustering[3] of the simulation decoys, which are further refined by the ModRefiner[4] and FG-MD[5] programs.

The major difference between QUARK and C-QUARK is that the sequence-based contact predictions, generated by NeBcon[6] and ResPRE (a new deep-learning based contact-map predictor, Li et al, in preparation), have been incorporated in the C-QUARK force filed to guide the folding simulations. The contact restraint potential is featured with a landscape of three gradients with continuous inflection slope at each gradient:

$$E_{cont}(d_{ij}) = \begin{cases} -U_{ij}, & d_{ij} < 8\text{Å} \\ -\frac{1}{2}U_{ij}\left[1 - sin\left(\frac{d_{ij}-(\frac{8+D}{2})}{d_b}\pi\right)\right], & 8\text{Å} \le d_{ij} < D \\ \frac{1}{2}U_{ij}\left[1 + sin\left(\frac{d_{ij}-(\frac{D+80}{2})}{(80-D)}\pi\right)\right], & D \le d_{ij} \le 80\text{Å} \\ U_{ij}, & d_{ij} > 80\text{Å} \end{cases} \qquad (1)$$

where d_{ij} is the $C\beta$-distance between the residue pair. The depth of the potential, U_{ij}, between residue pair (i and j) is proportional to the confidence score of the pair to be in contact. Overall, the 3-gradient potential is centered with a negative well at 8 Å cutoff, with a weak force in D (=8 Å + d_b) to 80 Å, followed by a stronger force in 8 Å to D, are introduced to push the target residue pairs towards the well when they are in a long distance. Both the height (U_{ij}) and the width (d_b) of the contact well are key parameters to determine the speed and the satisfactory rate of the contact map balanced with the inherent QUARK potential. Here, the height and width parameters, together with the weight and the number of contacts by different programs (ResPRE, NeBcon), are dependent on the length of the query sequence, the target type (trivial, easy, hard, and very hard) and the confidence score of different programs, which were systematically trained through a non-redundant set of 234 proteins (Mortuza et al, in preparation).

1. Xu, D.; Zhang, Y., Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins* 2012, **80**, 1715-35.
2. Xu, D.; Zhang, Y., Toward optimal fragment generations for ab initio protein structure assembly. *Proteins* 2013, **81**, 229-39.
3. Zhang, Y.; Skolnick, J., SPICKER: A clustering approach to identify near-native protein folds. *J Comput Chem* 2004, **25**, 865-71.
4. 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**, 2525-34.
5. Zhang, J.; Liang, Y.; Zhang, Y., Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling. *Structure* 2011, **19**, 1784-95.
6. He, B.; Mortuza, S. M.; Wang, Y.; Shen, H. B.; Zhang, Y., NeBcon: protein contact map prediction using neural network training coupled with naive Bayes classifiers. *Bioinformatics* 2017, **33**, 2296-2306.

Protein Structure Modeling by Predicted Distance instead of Contacts

Jinbo Xu

Toyota Technological Institute at Chicago, 6045 S Kenwood, IL, 60637, USA

jinboxu@gmail.com

Our CASP13 servers (RaptorX-Contact, RaptorX-DeepModeller and RaptorX-TBM) are built upon inter-residue distance prediction instead of contact prediction. We predict inter-residue distance using the deep learning method we developed for contact prediction [1]. The predicted distance is better than predicted contacts in the following aspects: 1) predicted distance contains finer-grained information than contacts; and 2) it is easier to (implicitly and explicitly) enforce physical constraints to our deep learning model when the goal is to predict a distance map instead of a contact map.

Methods

Predicting inter-residue distance. We use the same deep learning (DL) method described in Ref.[1] to predict inter-residue distance distribution for a query sequence. The only difference is that the goal in Ref.[1] is to predict the probability of two residues forming a contact while here we predict the distribution of the Euclidean distance between two residues. We discretize the inter-residue distance into 12 (or more) bins: <5Å, 5-6Å, …, 14-15Å, and >15Å. That is, 12 distance labels are used in our DL model, as opposed to 2 labels for contact prediction. The DL model for distance prediction is trained using the same training procedure, training set, and validation data as that for contact prediction. We also use the same input features, including sequential features (e.g., sequence profile and predicted secondary structure) and direct co-evolution information generated by CCMpred. Summing up the predicted probability values of the first 4 distance labels (corresponding to distance ≤8Å) and using the resultant summation as contact probability, our DL method for distance prediction has ~2% better contact prediction accuracy than our DL method for contact prediction (i.e., the model reported in Ref.[1]).

RaptorX-TBM. This is a new threading method described in our latest paper [2]. This new method employs predicted inter-residue distance to significantly improve sequence-template alignment and template selection. RaptorX-TBM works particularly well for a target with only remote templates because in this case it is hard to generate very accurate alignments and identify the best templates. Experimental results show that by using predicted distance, we can do much better than our previous threading method without using predicted distance. Finally, we generated the 3D models by MODELLER and Rosetta based upon the alignment generated by this threading method.

RaptorX-Contact. This is an ab initio folding method using predicted distance as restraints. No energy function is used. That is, we feed the predicted distance into CNS to reconstruct the 3D model of a target without using any template information. Predicted distance enables us to fold a protein much more accurately than by predicted contacts.

RaptorX-DeepModeller. It is an integration of RaptorX-TBM and RaptorX-Contact.

Results

1) Tested on the 37 CASP12 hard targets, RaptorX-Contact can generate correct folds (TMscore≥0.5) for around 20 of them. By contrast, the best CASP12 groups can generate correct folds for 11 of them.
2) For the results of RaptorX-TBM, please check out our latest paper [2] published by ISMB 2018 and Bioinformatics.
3) Tested on the 86 CASP12 domains, the 3D models generated by RaptorX-DeepModeller are about 10% better than RaptorX-TBM in terms of TMscore.

Availability: http://raptorx.uchicago.edu/

1. Wang S., Sun S., Li Z., Zhang R. and Xu, J. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLoS Computational Biology* **13**(1): e1005324.2017.
2. Zhu J., Wang S., Bu D. and Xu J. Protein threading using residue co-variation and deep learning. *Bioinformatics* 2018 (Proceedings of ISMB 2018).

End-to-End Deep Learning from raw Multiple Sequence Alignments to Contact Prediction

Claudio Mirabello and Björn Wallner

Linkӧping University, Dept. Physics, Chemistry, and Biology, Bioinformatics Division S-581 83 Linkӧping, Sweden

bjorn.wallner@liu.se

In the last few decades, huge efforts have been made in the bioinformatics community to develop machine learning-based methods for the prediction of structural features of proteins in the hope of answering fundamental questions about the way proteins function and about their involvement in several illnesses. The recent advent of Deep Learning has renewed the interest in neural networks, with dozens of methods being developed in the hope of taking advantage of these new architectures. On the other hand, most methods are still based on heavy pre-processing of the input data, as well as the extraction and integration of multiple hand-picked, manually designed features. Since Multiple Sequence Alignments (MSA) are almost always the main source of information in *de novo* prediction methods, it should be possible to develop Deep Networks to automatically refine the data and extract useful features from it. In this work, we propose a new paradigm for the prediction of protein structural features called rawMSA. The core idea behind rawMSA is borrowed from the field of natural language processing to map amino acid sequences into an adaptively learned continuous space. This allows the whole MSA to be input into a deep network, thus rendering sequence profiles, covariance analysis and other pre-calculated features obsolete. Further details are available at bioRxiv: https://doi.org/10.1101/394437

Methods

The rawMSA group participated in the TS and RR categories in CASP13. In the RR category we generate multiple contact maps from 12 deep network models. Ensembling is performed by averaging the softmax outputs of all models. The average output for class 1 (contact) is the final contact probability.

In the TS category we generated as many decoys as possible before the submission deadline with both CONFOLD[1] and the Rosetta Abinitio Relax protocol[2], depending on the target size and time constraints this resulted in 80 to 21,516 decoys per target (median: 1,298 decoys). In both cases, we selected the top contacts from the predicted contact map to constrain the folding procedure. We run each procedure many times with different contact selection thresholds (selecting from 0.1*L to 2*L top contacts). A filter to exclude obvious extended conformations was applied by requiring the “fatness” to be less than five. Fatness is defined as the ratio between the largest and smallest axis when representing the protein as an ellipsoid with the same moments of inertia. The remaining decoys were then scored and ranked by ProQ2[3] and the top five are submitted as TS models.

Results

We trained rawMSA on a large set of proteins and benchmarked it on 37 FM domains from CASP12 demonstrating that it performs on a par with the top ranked CASP12 methods in the inter-residue contact map prediction category, although no explicit correlated mutation or covariance information is calculated and used from the MSA. To ensure a fair comparison with the CASP12 predictors, we run the benchmark in the same conditions to which all the other predictors where subjected at the time of the CASP experiment (protein databases version, PDB version). We have also downloaded all the predictions made in CASP12 and evaluated them with the same system we used for our predictions. Results are shown in the Table 1.

Predictor	**Domain Count**	**L/5 LR Accuracy**
rawMSA CMAP	*37*	*43.8*
RaptorX-Contact	37	43.0
iFold_1	36	42.3
Deepfold-Contact	37	38.6
MetaPSICOV	37	38.4
MULTICOM-CLUSTER	37	37.9

Table 1: Comparison of rawMSA against the top 5 contact prediction methods in CASP12 using the L/5 Long-Range (LR) accuracy.

Availability

Datasets, network models and code to generate dataset and evaluate performance are available at: https://bitbucket.org/clami66/rawmsa

1. Adhikari, B., Bhattacharya. D., Cao, R., Cheng, J. CONFOLD: residue-residue contact-guided ab initio protein folding. *Proteins: Structure, Function, and Bioinformatics* **83**(8) (2015).
2. Raman, S., Vernon, R., Thompson, J., Tyka, M., Sadreyev, R., Pei, J., Kim, D., Kellogg, E., DiMaio, F., Lange, O. and Kinch, L. Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins: Structure, Function, and Bioinformatics*, *77*(S9) (2009).
3. Ray, A., Lindahl, E. & Wallner, B. Improved model quality assessment using ProQ2. *BMC Bioinformatics* **13,** 224 (2012).

Protein Structure Prediction by RBO Aleph in CASP13

M. Mabrouk[1], K. Stahl[1], S. Junghans[1], L. Hoenig[1] and O. Brock[1]

*1 - Robotics and Biology Laboratory, Technische Universit*ǟ*t Berlin*

oliver.brock@tu-berlin.de

RBO Aleph is a protein structure prediction server with a focus on free modeling targets. Our approach is based on two key ideas: 1) leveraging diverse information sources to gain knowledge, in the form of contacts, about the native conformation and 2) incorporating this knowledge into the energy landscape and using Model-based Search (MBS) to steer search towards low-energy regions. MBS builds a model of the energy landscape to focus sampling into low-energy regions. This approach enables us to efficiently exploit predicted residue-residue contacts in search.

The server is an update of previous versions that participated in CASP11 and CASP12. The pipeline logic has been rewritten using the Snakemake workflow management system[1], which provides a modular framework for quickly testing and replacing components in the pipeline. We identified domain boundary prediction, domain assembly, and model selection as major shortcomings in the previous version. We tackled the issue by replacing and adding new methods to these components. Last, we included an updated version of our contact prediction method (RBO-Epsilon), which was extended to use a fully convolutional network.

Methods

Pipeline overview. RBO Aleph[2] retrieves templates using a combination of scores from HHsearch[3], LOMETS[4], SparksX[5] and RaptorX[6]. In case templates are found, they are used to split the protein into domains, if not, the domain boundaries are predicted using two sequence-based domain prediction algorithms, PPRODO[7], DomPro[8] and DoBo[9]. Domains with available templates are modeled using Modeller and the top models are selected using QMEAN[10], a knowledge-based energy function. Free modeling domains are modeled using the method described below. Targets consisting of multiple domains are reconstructed by the domain assembly method AIDA[11].

Contact prediction. We use our own contact prediction method which was updated for CASP13 (server RBO-Epsilon). Our method extends over current approaches by combining evolutionary (GaussDCA[12], CCMpred[13], EVfold[14], GREMLIN[15], PSICOV[16]), sequence-based and physicochemical information (EPC-map[17]). We employ a deep fully convolutional network to effectively exploit the different profiles of the information sources and learn long-range dependencies between residue-residue contacts. A more detailed description of the methods can be found in the abstract of RBO-Epsilon in this issue.

Ab initio prediction. Similarly to CASP12, we use Model-based Search (MBS[18]) to leverage the predicted contacts in conformational search. MBS identifies funnels in the energy landscape and incrementally increases the sampling in the regions containing low-energy conformation. The predicted contacts are incorporated as distance constraints and added to the energy function to bias the search. Previous analysis[19] showed that MBS performs poorly compared to other methods when fed wrong contacts. This lies in its tendency to over-exploit the input information and thereby focus sampling on non-native space regions. In order to overcome this problem, we furthermore utilize Rosetta[20] Monte-Carlo-based search to sample the conformational space. Compared to MBS, Rosetta produces more diverse decoy sets. Combining both methods allows us to produce a decoy set that is diverse yet contains decoys which satisfy the majority of contacts. We finally use the QMEAN scoring function to select the top models from the decoys generated from MBS and Rosetta.

Availability

We offer access to our protein structure and contact prediction methods through a webserver under http://compbio.robotics.tu-berlin.de/rbo_aleph/

1. Köster, J., & Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, **28**(19), 2520-2522.
2. Mabrouk, M., Putz, I., Werner, T., Schneider, M., Neeb, M., Bartels, P., & Brock, O. (2015). RBO Aleph: leveraging novel information sources for protein structure prediction. *Nucleic acids research*, **43**(W1), W343-W348.
3. Söding, J., Biegert, A., & Lupas, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic acids research*, **33**(suppl 2), W244-W248.
4. Wu, S., & Zhang, Y. (2007). LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic acids research*, **35**(10), 3375-3382.
5. 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 query and corresponding native properties of templates. *Bioinformatics*, **27**(15), 2076-2082.
6. Peng, J., & Xu, J. (2011). RaptorX: exploiting structure information for protein alignment by statistical inference. Proteins: Structure, Function, and *Bioinformatics*, **79**(S10), 161-171.
7. Sim, J., Kim, S. Y., & Lee, J. (2005). PPRODO: prediction of protein domain boundaries using neural networks. Proteins: Structure, Function, and *Bioinformatics*, **59**(3), 627-632.
8. Cheng, J., Sweredoski, M. J., & Baldi, P. (2006). DOMpro: protein domain prediction using profiles, secondary structure, relative solvent accessibility, and recursive neural networks. *Data Mining and Knowledge Discovery*, **13**(1), 1-10.
9. Eickholt, J., Deng, X., & Cheng, J. (2011). DoBo: Protein domain boundary prediction by integrating evolutionary signals and machine learning. *BMC bioinformatics*, **12**(1), 43.
10. 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.
11. Xu, D., Jaroszewski, L., Li, Z., & Godzik, A. (2014). AIDA: ab initio domain assembly server. *Nucleic acids research*, **42**(W1), W308-W313.
12. Tetchner, S., Kosciolek, T., & Jones, D. T. (2014). Opportunities and limitations in applying coevolution-derived contacts to protein structure prediction. *Bio-Algorithms and Med-Systems*, **10**(4), 243-254.
13. Seemayer, S., Gruber, M., & Söding, J. (2014). CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. *Bioinformatics*, **30**(21), 3128-3130.
14. Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R., & Sander, C. (2011). Protein 3D structure computed from evolutionary sequence variation. *PloS one*, **6**(12), e28766.
15. Kamisetty, H., Ovchinnikov, S., & Baker, D. (2013). Assessing the utility of coevolution-based residue–residue contact predictions in a sequence-and structure-rich era. *Proceedings of the National Academy of Sciences*, **110**(39), 15674-15679.
16. 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**(2), 184-190.
17. Schneider, M., & Brock, O. (2014). Combining physicochemical and evolutionary information for protein contact prediction. *PloS one*, **9**(10), e108438.
18. Brunette, T. J., & Brock, O. (2005). Improving protein structure prediction with model-based search. *Bioinformatics*, **21**(suppl 1), i66-i74.
19. Mabrouk, M., Werner, T., Schneider, M., Putz, I., & Brock, O. (2016). Analysis of free modeling predictions by RBO aleph in CASP 11. *Proteins: Structure, Function, and Bioinformatics*, **84**, 87-104.
20. Rohl, C. A., Strauss, C. E., Misura, K. M., & Baker, D. (2004). Protein structure prediction using Rosetta. *Methods in enzymology*, **383**, 66-93.

Contact Prediction by RBO-Epsilon v2 in CASP13

K. Stahl[1], S. Junghans[1], M. Mabrouk[1] and O. Brock[1]

[1] - Robotics and Biology Laboratory, Technische Universität Berlin

oliver.brock@tu-berlin.de

To predict contacts with high accuracy, it is vital to leverage as much and diverse information as possible. RBO-Epsilon therefore combines evolutionary, sequence-based, and physicochemical information. These sources of information are complementary. By combining them effectively, we can compensate the shortcomings of one type based on the strength of another. Our approach for this is based on deep learning and utilizes stacking. Stacking treats the combination process as a learning problem. With the help of indicator features we learn to leverage the most effective source of information. We also simplified the feature set conventionally used so that we can learn on more data and increase model complexity. EPSILON-CP v1[1] ranked 5th in the final evaluation of CASP 12 [2]. We extend the original EPSILON-CP to predict the complete contact map at once using a fully convolutional neural network with 68 hidden layers. The depth in addition to dilated convolutions allows the network to learn long-range dependencies between residue-residue contacts improving contact prediction significantly[3].

Methods

RBO-Epsilon combines evolutionary, sequence-based, and physicochemical information. The physicochemical information stem from EPC-map[4], which ranked amongst the top contact predictors in CASP11. We use the sequence-based feature set employed by MetaPSICOV[5] stage1 and RaptorX[3], that include the amino acid composition, secondary structure prediction, solvent accessibility and column entropy amongst other features. Building on the idea of PconsC[6] and MetaPSICOV to include multiple different co-evolutionary information, we extend the feature set to include the prediction of GaussDCA[7], in addition to CCMpred[8], FreeContact[9].

The feature set is initially high dimensional with 672 features. Using a feature importance analysis, the dimensionality (including the newly added features) is reduced to 171 features in v1 and further to 35 in v2, enabling the use of a more complex neural network. The feature importance is computed by XGBoost[10] (eXtreme gradient boosting), a decision tree-based approach. XGBoost partitions the dataset based on features that best separates the classes (here contacts and non-contacts). Features that are higher up in the tree are deemed more important. The final feature set is a mix of high level features (EPC-map prediction, co-evolutionary information) and crude sequence-based features which may also act as indicator variables for the more high-level features. It can therefore be seen as a variant of stacking.

The final model is trained on 7480 proteins. The network is implemented in PyTorch[11]

Availability

RBO-EPSILON is available as a web server: https://compbio.robotics.tu-berlin.de/epsilon The code is available from the author on request.

1. Stahl, K., Schneider, M., & Brock, O. (2017). EPSILON-CP: using deep learning to combine information from multiple sources for protein contact prediction. *BMC bioinformatics*, ***18***(1), 303.
2. Schaarschmidt, J., Monastyrskyy, B., Kryshtafovych, A., & Bonvin, A. M. (2018). Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age. *Proteins: Structure, Function, and Bioinformatics*, **86**, 51-66.
3. Wang, S., Sun, S., Li, Z., Zhang, R., & Xu, J. (2017). Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS computational biology*, **13**(1), e1005324.

4. Schneider, M., & Brock, O. (2014). Combining physicochemical and evolutionary information for protein contact prediction. *PloS one*, **9**(10), e108438.
5. Jones, D. T., Singh, T., Kosciolek, T., & Tetchner, S. (2014). MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics*, **31**(7), 999-1006.
6. Skwark, M. J., Abdel-Rehim, A., & Elofsson, A. (2013). PconsC: combination of direct information methods and alignments improves contact prediction. *Bioinformatics*, **29**(14), 1815-1816.
7. Baldassi, C., Zamparo, M., Feinauer, C., Procaccini, A., Zecchina, R., Weigt, M., & Pagnani, A. (2014). Fast and accurate multivariate Gaussian modeling of protein families: predicting residue contacts and protein-interaction partners. *PloS one*, **9**(3), e92721.
8. Seemayer, S., Gruber, M., & Söding, J. (2014). CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. *Bioinformatics*, **30**(21), 3128-3130.
9. Kaján, L., Hopf, T. A., Kalaš, M., Marks, D. S., & Rost, B. (2014). FreeContact: fast and free software for protein contact prediction from residue co-evolution. *BMC bioinformatics*, **15**(1), 85.
10. Chen, T., He, T., & Benesty, M. (2016). xgboost: Extreme gradient boosting. R package version 0.4–4.
11. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., ... & Lerer, A. (2017). Automatic differentiation in pytorch.

Contact Prediction by Stacked Fully Convolutional Residual Neural Network Using Coevolution Features from Deep Multiple Sequence Alignments

Yang Li[1,2], Chengxin Zhang[2], Dongjun Yu[1] and Yang Zhang[2]

1 School of Computer Science and Engineering, Nanjing University of Science and Technology, Xiaolingwei 200, Nanjing, China, 210094; 2Department of Computational Medicine and Bioinformatics, Department of Biological Chemistry, University of Michigan, 100 Washtenaw Ave, Ann Arbor, MI 48109

yangzhanglab@umich.edu

We developed the ResTriplet protein contact prediction server. It combined three different evolutionary coupling approaches by a set of fully convolutional residual neural networks.

Methods

For a query sequence, a deep multiple sequence alignment (MSA) was built by HHblits[1] search of UniClust30[2] database, followed by jackhmmer[3] search through UniRef90[4]. Instead of being directly concatenated to HHblits MSA, the jackhmmer hits were converted into an HHblits format sequence database against which a second HHblits search was performed. The combined MSA from the two HHblits runs was further enriched by hmmsearch[3] through MetaClust[5] database.

From this deep MSA, we calculated evolutionary couplings using three different approaches: inverse covariance estimation from the MSA represented in one-hot-encoding (with Tikhonov regularization instead of the L1 regularization in previous work[6]), pseudo-likelihood maximization,[7] and covariance calculation. Each approach provided a (21**L*)*(21**L*) evolutionary coupling matrix for a protein sequence with *L* residues and 21 residue types (20 kinds of amino acids plus gap).

Instead of summing up the (21**L*)*(21**L*) evolutionary coupling matrices into *L***L* contact matrices,[6,7] the values of these three matrices were directly used as input features for a stack of residual convolutional neural networks (CNNs). The stacking approach started with three separate CNNs, each of which was trained separately on one of the three evolutionary matrices as input features and had 46 layers. The predicted contacts of these three CNNs, as well as predicted secondary structures[8] derived from the same deep MSA, were used as the input features for one last CNN with 6 layers to generate the final contact prediction output. Apart from secondary structure, the CNN models did not use any other linear features such as sequence profile or solvent accessibility.

If the query sequence was predicted by ThreaDom[9] as a multi-domain protein, contact predictions for both the full length protein and the individual domains are separately performed. Meanwhile, the top five LOMETS[10] templates were used to generate a homology model by MODELLER[11]. If contact agreement between the MODELLER model and ResTriplet prediction for an individual domain was higher than agreement for the corresponding domain region in the full length ResTriplet prediction, contacts for this region of the full length prediction would be replaced by prediction for the individual domain. It should be noted that we used the same ResTriplet algorithm for contact prediction using full length sequence and that using individual domain sequence. Nor did we attempt to re-normalize contact confidence scores predicted for full length sequence and those predicted for domain sequence.

To test the effect of different ensemble approaches, we constructed a separate CASP13 server TripletRes. While ResTriplet used stacking to combine the three individually trained CNNs by another CNN, TripletRes trained all four CNNs together in an end-to-end fashion.

Availability

The on-line ResTriplet server and its CASP13 the prediction results are available at https://zhanglab.ccmb.med.umich.edu/ResTriplet/.

1. Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature methods 2012;9(2):173-175.
2. Mirdita M, von den Driesch L, Galiez C, Martin MJ, Soding J, Steinegger M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Res 2017;45(D1):D170-D176.
3. Eddy SR. Accelerated Profile HMM Searches. Plos Comput Biol 2011;7(10).
4. Suzek BE, Wang YQ, Huang HZ, McGarvey PB, Wu CH, Consortium U. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 2015;31(6):926-932.
5. Steinegger M, Soding J. Clustering huge protein sequence sets in linear time. Nat Commun 2018;9.
6. Jones DT, Buchan DW, Cozzetto D, Pontil M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 2012;28(2):184-190.
7. Seemayer S, Gruber M, Söding J. CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. Bioinformatics 2014;30(21):3128-3130.
8. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. Journal of Molecular Biology 1999;292(2):195-202.
9. Xue Z, Xu D, Wang Y, Zhang Y. ThreaDom: extracting protein domain boundary information from multiple threading alignments. Bioinformatics 2013;29(13):i247-i256.
10. Wu ST, Zhang Y. LOMETS: A local meta-threading-server for protein structure prediction. Nucleic Acids Res 2007;35(10):3375-3382.
11. Šali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. Journal of molecular biology 1993;234(3):779-815.

Protein model quality estimation by single model-based method SART in CASP13

Kun-Sop Han and Myong-Ho Choe

Department of Life Science, University of Science, Unjong-District, Pyongyang, DPR Korea

hks1981@star-co.net.kp

Quality Assessment (QA) of protein models is an essential component in any protein structure prediction method and is important for determining its usefulness for specific application. We participated in QA category of CASP13 with two prediction methods. The method described here is labeled as group "SASHAN", number 220. This method is a new single model-based quality assessment program that predicts local as well as global quality of protein models. A new single model-based global quality score SART_G is a linear combination of 10 components (including agreements in secondary structure, solvent accessibility and contact, and various statistical potentials) extracted in model level from protein model of interest. For global score, most of components are normalized by the target sequence length to enable comparisons between proteins. A new single model-based local quality score SART_L is based on linear combination of 9 components (similar to SART_G) extracted from a sphere centered on the residue of interest. For local score, after the true distance, d, is converted to the S-score with distance threshold d_0=3.8Å, $S=1/(1+(d/d_0)^2)$, linear regression analysis is performed between 9 components and S-scores. The per residue predicted distance deviation SART_L is calculated by SART_L=$d_0(1/\text{S-score}-1)^{1/2}$. We put all SART_L>15Å to 15Å. 34337 CASP9 models are used as training set.

Methods

1. SART_G: Single model-based global quality score

We extract 10 components from a protein model, and then combine those components to obtain the single model-based global quality score SART_G.

- Components of SART_G

SS_8BIN: The number of residues that the predicted secondary structure (by SSpro8_5.1 of SCRATCH[1]) equals to the calculated secondary structure (by DSSP[2]) and the predicted solvent accessibility (by ACCpro_5.1 of SCRATCH) equals to the transformed binary solvent accessibility (by DSSP) is divided by L (the sequence length of target protein).

SS_8: The 8 states-agreement number between the predicted secondary structure of target and the calculated secondary structure of model is divided by L.

ACC_{SPE}: Solvent accessibility of target is predicted by ACCpro20_5.1 of SCRATCH. The calculated solvent accessibilities of the model are divided by the maximum solvent accessibility of the corresponding residue. The Spearman correlation coefficient (R_{SPE}) between the divided solvent accessibilities of target and model is calculated. R_{SPE} is multiplied by the residue fraction of the model.

SS_ResiEplus: Only 12 residues are considered. The sum of the secondary structure-specific residue pair potentials (bigger than zero) of all possible 12 residue pairs with the distance of 3-25 Å in the model is divided by $L^{1.3}$.

ACC_ResiE: The sum of solvent accessibility-specific residue pair potentials of all possible residue pairs with the distance of 3-25 Å in the model is divided by $L^{1.2}$.

TorE: The torsion potential is derived on the basis of new division of Ramachandran plot into 8 regions. The sum of torsion potentials of all tripeptides in the model is divided by L.

TotActplus: The number of tripeptides in the model which has the torsion potential bigger than zero is divided by L.

Hydrophobic_Ratio: The number of 8 hydrophobic residues in the model which are in buried state is divided by the total number of 8 hydrophobic residues in the model.

BE_Potential: BE_Potential is a component of indicating how well the burial propensities of 20 amino acids derived from 3739 chains are represented in the model.

RRcon: The number of residue pairs in the model which has the contact distance less than 8Å and belongs to the top 2×L residue pairs with the highest contact probability by the in-house residue-residue contact prediction program is divided by 2 × L, obtaining RRcon.

- Construction of SART_G

SART_G is a linear combination of 10 components described above. Weights of 10 components and constant term are obtained by linear regression analysis between 10 components and GDT_TS scores of 34337 CASP9 models.

2. SART_L: Single model-based local quality score

We extract 9 components from the sphere (radius 12 Å) centered on residue of interest, and then combine those components to obtain the single model-based local quality score SART_L.

- Components of SART_L

R_SS_8 is the 8 states-agreement number between the predicted and calculated secondary structure of amino acids within the sphere.

R_ACC_{BIN} is the 2 states-agreement number between the predicted and the transformed 2 states-solvent accessibility of amino acids within the sphere.

R_ACC_{AVE} is the arithmetic mean of solvent accessibilities (divided by the corresponding maximum solvent accessibility) of the amino acids within the sphere.

R_ACC_{PEA} = 1 - R_{PEA}. R_{PEA} is the Pearson correlation coefficient between the predicted and the calculated decimal solvent accessibilities of amino acids within the sphere.

R_SS_8BIN is the number of the residues within the sphere that the predicted secondary structure equals to the calculated secondary structure and the predicted solvent accessibility equals to the transformed binary solvent accessibility.

R_PlusBuried_Count is the number of the residues within the sphere that have the burial propensity bigger than zero and are in the buried status.

R_RRcon is the number of residue pairs within the sphere which has the contact distance less than 8Å and belongs to the top 3 × L residue pairs with the highest contact probability by the in-house residue-residue contact prediction program.

R_TotEplus is the sum of the torsion potentials bigger than zero within the sphere.

R_TorActplus is the number of tripeptides within the sphere which has the torsion potential bigger than zero.

- Construction of SART_L

The SART_L is based on linear combination of 9 components described above. The true distance, d, is converted to the S-score with distance threshold d_0=3.8Å, $S = 1 / (1 + (d / d_0)^2)$. Weights of 9 components and constant term are obtained by linear regression analysis between 9 components and S-score calculated from 6818635 residues of 34337 CASP9 models. The per residue predicted distance deviation SART_L is calculated by SART_L = $d_0 (1 / \text{S-score} - 1)^{1/2}$. We put all SART_L>15Å to 15Å.

Results

We use 3 metrics (i.e. average per target Pearson correlation coefficient, average per target quality loss and AUC (cutoff = 0.5 GDT_TS)) to assess the global prediction methods for 91 CASP11 targets.

Table 1. Comparison of SART_G with single model-based global QA methods in CASP11

Methods	stage1			stage2		
	Average Pearson	Average quality loss	AUC	Average Pearson	Average quality loss	AUC
SART_G	0.663	0.075	0.934	0.398	0.078	0.925
363(Wang-SVM)	0.646	0.107	0.905	0.354	0.094	0.861
132(ProQ2refine)	0.644	0.091	0.929	0.360	0.069	0.914
420(MULTICOM_cluster)	0.640	0.099	0.921	0.394	0.075	0.905
338(ProQ2)	0.638	0.087	0.924	0.364	0.063	0.909

We use 2 metrics (i.e. average per model Pearson correlation coefficient and MCC (cutoff = 3.8Å)) to assess the local prediction methods for 63 CASP11 targets.

Table 2. Comparison of SART_L with single model-based local QA methods in CASP11

Methods	stage1		stage2	
	Average Pearson	MCC	Average Pearson	MCC
338(ProQ2)	0.351	0.537	0.475	0.530
132(ProQ2refine)	0.354	0.537	0.476	0.526
SART_L	0.295	0.540	0.392	0.519
083(Wang_deep_3)	0.262	0.508	0.334	0.465
020(Wang_deep_1)	0.227	0.498	0.301	0.460
031(Wang_deep_2)	0.268	0.497	0.339	0.457
363(Wang_SVM)	0.266	0.442	0.312	0.396

Availability

Manuscript for SART is in preparation.

1. Cheng,J., et al. (2005) SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Res. **33**, W72–W76.
2. Kabsch,W., Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers, **22**(12), 2577-2637.
3. Kryshtafovych,A., Barbato,A., Fidelis,K., Monastyrskyy,B., Schwede,T., and Tramontano,A. (2014) Assessment of the assessment: Evaluation of the model quality estimates in CASP10. Proteins, **82**, 112-126.
4. Liu,T., Wang,Y., Eickholt,J. & Wang,Z. (2016) Benchmarking Deep Networks for Predicting Residue-Specific Quality of Individual Protein Models in CASP11. Scientific Report, 6, 1930.

RADI: Protein contact predictions using a reduced alphabet and direct-coupling analysis

B. Anton[1], M. Besalú[2], J. Bonet[3], G De las Cuevas[4], O. Fornes[5], N. Fernandez-Fuentes[6,7]and B. Oliva[1]

1 - Structural Bioinformatics Lab (GRIB-IMIM), Department of Experimental and Health Science, University Pompeu Fabra, Catalonia, Spain. 2 –Departament de Matematiques I Informatica, Universitat de Barcelona, Catalonia, Spain.3 - Laboratory of Protein Design & Immunoengineering, School of Engineering, Ecole Polytechnique Federale de Lausanne, Switzerland.4 –Institue für Theoritische Physik, School of Mathematics, Computer Sciences and Physic, Universität Innsbruck. Austria. 5 - Centre for Molecular Medicine and Therapeutics, Department of Medical Genetics, BC Children's Hospital Research Institute, University of British Columbia, Canada. 6 - Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, UK7 –Universitat of Vic - Universitat Central de Catalunya, Catalonia, Spain.

bernat.anton@upf.edu, mbesalu@ub.cat, jaume.bonet@epfl.ch, gemma.delascuevas@uibk.ac.at,
oriol@cmmt.ubc.ca, naf4@aber.ac.uk, narcis.fernandez@uvic.cat, boliva@upf.edu

The identification of coevolved residue pairs in protein sequences is widely used to help the prediction of three-dimensional (3D) structure in proteins[1]. Besides functional implication, often pairs of coevolved residues inform of the 3D closeness and thus it can be used to guide structural prediction of proteins in the form of distance restraints[2]. Direct-coupling analysis (DCA) is used currently to identify such pairs of residues but at a high computational cost[3]. We recently developed a novel computational approach named RADI, for Reduced Alphabet Direct Information[4] which present novel ideas to improve the speed of calculation of direct information values. By using a simplified alphabet, i.e. grouping amino acids with similar physicochemical properties, RADI achieved can achieved a reduction of the computational without loss of accuracy as proved on a benchmark set. We have now applied RADI on a blind test using the sequences submitted to CASP13 under residue-residue contact prediction section. Overall, we provided prediction for 66 submitted targets.

Methods

The protocol followed to computed DI values from RADI as follow:

(i) Generation of multiple-sequence alignments (MSAs)

MSAs were created using the script "buildmsa.py" included in the RADI Git repository. First, the script builds a profile of the query searching for similar sequences in the uniref50 database with MMseqs2[5]. Next, it uses the query profile to find more sequence relatives in the uniref100 database. Then, the script builds a MSA of the query and the identified sequences (up to 100,000) with FAMSA[6]. Finally, it removes the columns of the MSA with insertions in the query. Note that MMseqs2 is executed with options "-s 7.5" and "--max-seq-id 1.0" for a more sensitive search.

(ii) Secondary structure prediction

The secondary structures were predicted using SABLE[7] and a 3-state alphabet, namely: helix (H), beta(E) and coil (C).

(iii) Calculation of DI values

The calculation of DI values was done using the original DCA algorithm as implemented in RADI utilizing four different alphabets, namely RA0, RA1, RA2, and RA3 (for more information on the method please refer to original publication[4].

(i) RA0 stand for an alphabet of size q = 21 (i.e. 20 different amino acids plus the gap)

(ii) RA1 has a q = 9 represented by Positively charged: {Arg, His, Lys}. Negatively charged: {Asp, Glu}. Polars: {Ser, Thr, Asn, Gln}. Aliphatics: {Ala, Ile, Leu, Met, Val}. Aromatics: {Phe, Trp, Tyr}. Single groups: {Cys}, {Gly}, {Pro} and the gap;

(iii) RA2 has a q = 5 represented by Polar: {Arg, His, Lys, Asp, Glu, Ser, Thr, Asn, Gln, Cys}. Non-polar: {Ala, Ile, Leu, Met, Val, Phe, Trp, Tyr}. Single groups: {Gly}, {Pro} and the gap; and

(iv) RA3 has a q = 3 represented by Polar: {Arg, His, Lys, Asp, Glu, Ser, Thr, Asn, Gln, Cys, Gly}. Non-polar: {Ala, Ile, Leu, Met, Val, Phe, Trp, Tyr, Pro}. Single groups: gap

For each of the alphabet, i.e. RA{0-3} DI values are acquire for pair of amino acid belonging to two different secondary structures, i.e. pairs of residues within same secondary structure were not considered.

(iv) Selection and submission of top DI values

The top 40 DI value on each calculation, i.e. RA{0-3} were considered. In the case of shorter proteins/domains, i.e. smaller than 60 residues, the number of predictions considered was only the top 20 DI values.

Availability

RADI is available at: https://github.com/structuralbioinformatics/RADI

1. Marks, D.S., Colwell, L.J., Sheridan R., Hopf T.A., Pagnani A., Zecchina R., Sander, C. (2011). Protein 3D structure computed from evolutionary sequence variation. *PLoS One*. **6**, e28766.
2. Ovchinnikov, S., Kinch, L., Park, H., Liao, Y., Pei, J., Kim, D.E., Kamisetty, H., Grishin, N.V., Baker, D. (2015) Large-scale determination of previously unsolved protein structures using evolutionary information. *Elife*. **4**, e09248.
3. 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. *Proc Natl Acad Sci U S A*. **108**, 1293-1301.
4. Anton, B., Besalu, M., Fornes, O., Bonet, J., Cuevas G. De las, Fernandez-Fuentes N., Oliva, B. (2018) RADI (Reduced Alphabet Direct Information): Improving execution time for direct-coupling analysis. *bioRxiv*. **406603**, doi: https://doi.org/10.1101/406603
5. Steinegger, M., Soding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol*. **35**, 1026-1028.
6. Deorowicz, S., Debudaj-Grabysz, Gudys, A. (2016) FAMSA: Fast and accurate multiple sequence alignment of huge protein families. *Sci Rep*. **6**, 33964.
7. Adamczak, R., Porollo, A., Meller, J. (2005). Combining prediction of secondary structure and solvent accessibility in proteins. *Proteins*. **59**, 467-475

Smooth orientation-dependent scoring function for coarse-grained protein quality assessment

S. Grudinin[1] and M. Karasikov[2]

1 - Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France, 2 - Department of Computer Science, ETH Zurich, Zurich, 8092, Switzerland

Sergei.Grudinin@inria.fr

Protein quality assessment (QA) is a crucial element of protein structure prediction, a fundamental and yet open problem in structural bioinformatics. QA aims at ranking predicted protein models to select the best candidates. Although consensus-model QA methods often outperform single- model QA methods, their performance substantially depends on the pool of available candidates. This makes single-model QA methods a particularly important research target since these usually assist when sampling the candidate models.

Methods

We present a novel single-model QA method called SBROD[1]. The SBROD (Smooth Backbone- Reliant Orientation-Dependent) method uses only the backbone protein conformation, and hence it can be applied to scoring coarse-grained protein models. SBROD deduces its scoring function from a training set of protein models (server submissions from previous CASP rounds). The SBROD scoring function is composed of four terms related to different structural features. These are relative residue-residue orientations, contacts between backbone atoms, hydrogen bonds and solvent-solvate interactions. The model is then trained using linear ridge regression to predict the GDT-TS score of the models in the training set. The obtained scoring function is smooth with respect to atomic coordinates and thus is potentially applicable to continuous gradient-based optimization of protein conformations. Furthermore, it can also be used for coarse-grained protein modeling and computational protein design.

Results

We evaluated SBROD on diverse datasets (CASP11, CASP12, and MOULDER) and proved that it achieves the state-of- the-art performance among single-model QA methods. In the CASP13 exercise, we applied SBROD to the QA category of targets using two server and one human group. Each group was running a SBROD model trained on slightly different datasets (server CASP submissions for rounds 5-11 and 5-12, and both server and human submissions for rounds 5-12).

Availability

The standalone application implemented in C++ and Python is freely available at https://gitlab.inria.fr/grudinin/sbrod and supported on Linux, MacOS, and Windows.

1. Karasikov,M., Pages,G., & Grudinin,S. Smooth orientation-dependent scoring function for coarse-grained protein quality assessment. *In revision.*

Scoring protein models with a hybrid back-propagation Levenberg-Marquardt artificial neural network

Eshel Faraggi[1] and Andrzej Kloczkowski[2]

1- IUPUI Physics, 2- NCH Math Med

efaraggi@gmail.com

The advent of larger and larger datasets, coupled with the complexity of characterizing protein models, represent a challenge to artificial neural networks. More input features describing the protein, and the abundance of training models, lead to highly dimensional input feature space with an abundance of local minima and a limited ability to sample the parameter space. To address this we propose a hybrid of the back-propagation algorithm, where a subset of the weight space undergoes a Levenberg-Marquardt step at each epoch. This approach is useful in cases where a crucial part of the network is of a limited extent, while other parts may be very large. This is the case for this problem, where the desired output is a single number representing the closeness of the model to a native structure. In addition, we describe introducing associative memory and additional input features into Seder.

Molecular Dynamics Based Protein Structure Refinement via Solvation Force Fluctuation Enhancement

Lianqing, Zheng[1], .Dongsheng Wu[1], and Wei Yang[1,2*]

1Institute of Molecular Biophysics, Florida State University, Tallahassee, 32306, 2Department of Chemistry and Biochemistry, Florida State University, Tallahassee, 32306

*yyang2@fsu.edu

In CASP 12, a preliminary enhanced sampling molecular dynamics method was employed for protein structure refinement. Specifically, solvation force, e.g. protein-water interaction, was utilized as the order parameter for sampling enhancement treatment and its various orders of environment responses were sampled via a high-order orthogonal space sampling strategy. The development of this method was motivated by the mechanism underlying chaperon catalyzed protein mis-folding/folding transition processes. In the past two years, this solvation force orthogonal space sampling method was further enriched and applied for protein structure refinement in CASP13. With various new technical treatments, improved sampling was obtained for aggressive refinement where no restraint is imposed, and no special selection is applied.

Methods

In certain aspect, protein structure refinement can be viewed as the process enabling protein mis-folding/folding transitions, which can take much longer time than natural thermal unfolding/folding transitions. Correspondingly, it is impractical for brute-force molecular dynamics simulation based methods to enable robust high-accuracy protein structure refinement for the fact that long-timescale thermal mis-folding/unfolding (or partial mis-folding/unfolding) transitions usually need to occur before the refinement enters the productive refolding stage. Therefore, most molecular dynamics simulation refinements have been performed conservatively under restraint treatments in order to gain consistent score increases but with limited performance in generating high-accuracy structures.

Chaperone proteins allow transitions from misfolding to folding states to occur in an accelerative and directional manner through repetitive changes of water activities surrounding the protein. Inside Chaperone, misfolding/folding transitions occur through mechanical work pathways, where usually no large scale thermal unfolding is needed for the completion of the refolding process. Motivated by this, in our CASP journey, we designed a specialized enhanced sampling strategy on one hand to enable aggressive protein structure refinement and on the other hand to understand detailed processes of chaperone induced misfolding/folding transitions. Specifically, solvation force, e.g. protein-water interaction, is employed as the order parameter for enhanced sampling treatment; thereby like being inside Chaperone, repetitive dry-to-wet solvation phase transitions can occur around a misfolded protein and such solvation force fluctuation can generate large work to catalyze the refolding process. As is known, each dry/wet phase transition in Chaperone may take millisecond, during which solvation force coupled protein structure changes can sufficiently occur. To speed up the sampling of protein structure responses, we take advantage of our orthogonal space sampling strategy[1,2].

In practice, our protein structure refinement was performed based on explicitly solvated (with the modified TIP3) all-atom CHARMM36* model. No restraint was imposed and no structure selection or other post-processing was made.

Results

In CASP12, the solvation force based sampling strategy was first time sysmatically used. As discussed by the evaluator, this method led to several notable high-accuracy refinement, however on average obtained structures were worse than initial models. After analyzing these results (unpublished), we realized that sampling on solvation force coupled protein structure transitions was too moderate; herein, solvation force fluctuation induced work was insufficiently applied and led to often inadequate time for the protein to move into the productive refolding stage.

Before CASP13, we faced two options, either adding restraints to perform conservative refinement to improve overall score but losing our hope of reaching ultimate targets or further speeding up sampling to hopefully gain sufficient time to drive the protein into the refolding stage. We chose the latter approach; specifically we updated the method to the fifth-order scheme and changed the tempering treatment that allows for 1500 K effective temperature speedup of orthogonal space coupling. As revealed by our own observations, improved sampling was enabled in our CASP 13 exercises. In our presentation, how improved sampling was translated into refinement improvement will be discussed.

Availability

Our detailed molecular dynamics trajectories and assorted results are available upon request after the summary paper is published.

1. Zheng, L., Chen, M. & Yang, W. (2008). Random walk in orthogonal space to achieve efficient free energy simulation of complex systems. *Proc. Natl. Acad. Sci.* **105**, 20227-20232.
2. Zheng, L. & Yang, W. (2012) Practically efficient and robust free energy calculations: Double-integration orthogonal space tempering. *J. Chem. Theor. Comput.* ***8***, 810–823.

Prediction of Protein Complex Structures by GALAXY in CASP13

Taeyong Park, Minkyung Baek, Hyeonuk Woo, and Chaok Seok

Department of Chemistry, Seoul National University, Seoul 151-747, Republic of Korea

chaok@snu.ac.kr

Seok-assembly is an automated homo- and hetero-oligomer structure prediction server. Human predictions for oligomer targets generated by running GALAXY programs manually were submitted by the human group Seok.

Methods

The overall pipeline for the Seok-assembly server is presented in **Figure 1**. GalaxyHomomer[1] which performs both template-based and *ab initio* homo-oligomer structure prediction was used to predict the structures of homo-oligomer targets. An improved version of GalaxyHomomer that incorporates a new *ab initio* docking program for oligomers of C_n symmetry GalaxyTongDock_C with an improved scoring function was used. To predict the structures of hetero-oligomer targets, structure of each subunit was first predicted by Seok TS protocol for monomer subunit or by GalaxyHomomer for homo-oligomer subunit. The predicted subunit structures were docked to form complex structures by using an in-house *ab initio* protein-protein docking program GalaxyTongDock. GalaxyTongDock performs FFT-based, low-resolution protein-protein docking and selects docking poses after clustering. Ten complex structures generated by GalaxyTongDock underwent optimization considering structure flexibility using GalaxyRefineComplex[2], resulting in five final models. In the case of H1021 ($A_6B_6C_6$) and H1022 (A_6B_3), symmetric axes of homo-oligomer subunits were matched during docking. The human predictions basically followed the overall server pipeline except that information from literature search and human insights were used in the stages of template selection, restraint generation, and model selection. In addition, information about protein interface was utilized via interface and/or block options during *ab initio* docking.

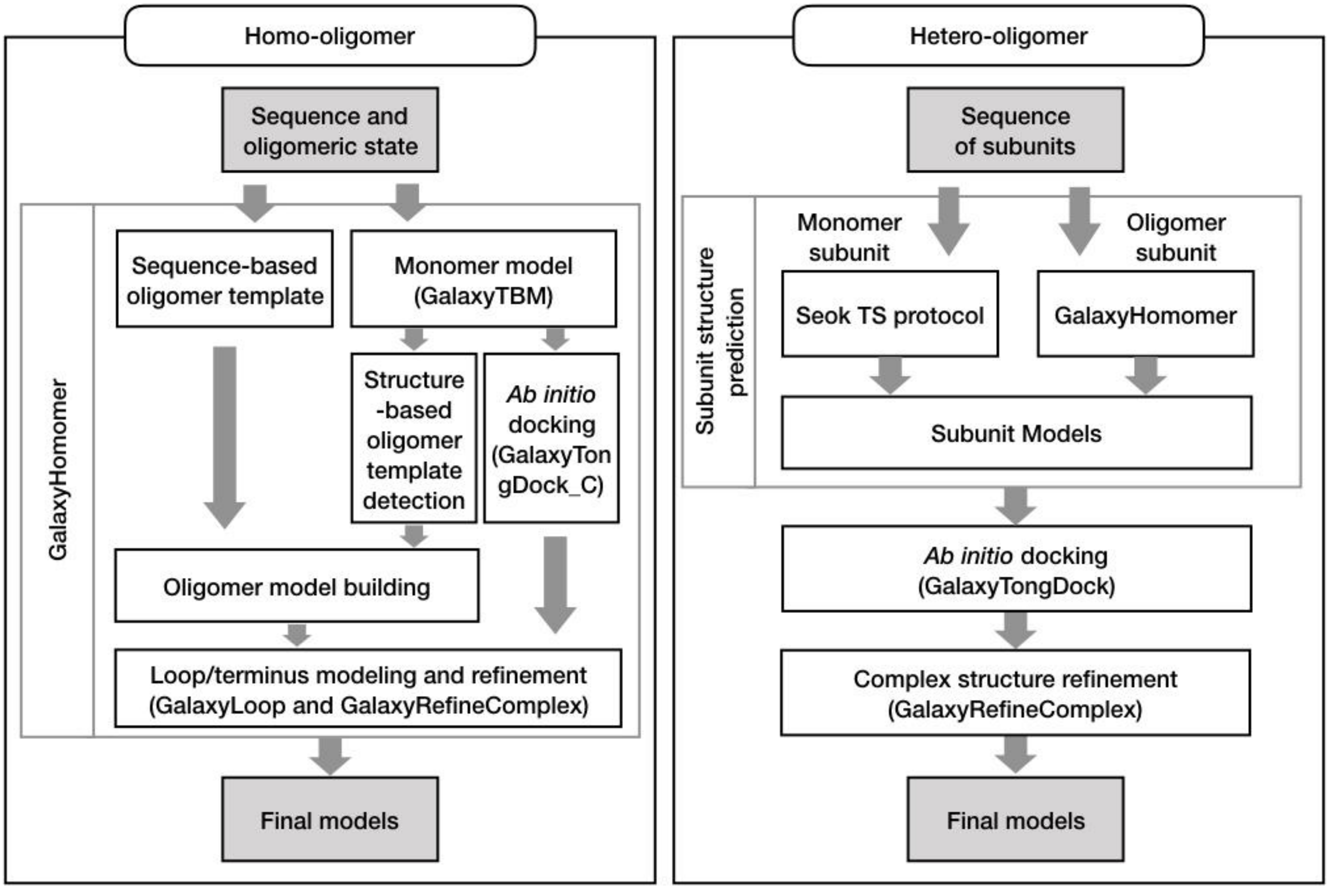

Homo-oligomer
Sequence and oligomeric state
GalaxyHomomer
Sequence-based oligomer template
Monomer model (GalaxyTBM)
Structure -based oligomer template detection
Ab initio docking (GalaxyTongDock_C)
Oligomer model building
Loop/terminus modeling and refinement (GalaxyLoop and GalaxyRefineComplex)
Final models
Hetero-oligomer
Sequence of subunits
Subunit structure prediction
Monomer subunit
Oligomer subunit
Seok TS protocol
GalaxyHomomer
Subunit Models
Ab initio docking (GalaxyTongDock)
Complex structure refinement (GalaxyRefineComplex)
Final models

Figure 1. Seok-assembly pipeline for CASP13

Availability

The GALAXY programs used in Seok-assembly are available on the GalaxyWEB web page at http://galaxy.seoklab.org.

1. Baek,M., Park,T., Heo,L., Park,C., Seok,C. GalaxyHomomer: a web server for protein homo-oligomer structure prediction from a monomer sequence or structure. Nucleic Acids Res. 2017;45:W320-W4.
2. Heo,L., Lee,H., Seok,C. GalaxyRefineComplex: Refinement of protein-protein complex model structures driven by interface repacking. Sci Rep. 2016;6:32153.

GALAXY in CASP13: Automated Protein Tertiary Structure Prediction

Minkyung Baek[1], Jonghun Won[1], Sohee Kwon[1], Jinsol Yang[1], Sangwoo Park[1], Taeyong Park[1], Hyeonuk Woo[1], Beomchang Kang[1], Jungun Park[1], Martin Steinegger[1,2], and Chaok Seok[1]

1- Department of Chemistry, Seoul National University, Seoul 08826, Republic of Korea, 2- Quantitative and Computational Biology group, Max-Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany

chaok@snu.ac.kr

Seok-server performed fully automated template-based protein tertiary structure predictions[1] for TS targets. A meta-server Seok-refine submitted predictions for TS targets by refining structures selected among the CASP server models. A simplified version of GalaxyRefine-CASP13[2] was used for model refinement.

Methods

Seok-server protocol for TS targets using improved GalaxyTBM

The protein tertiary structure prediction pipeline of Seok-server consists of the following steps: (1) modeling unit detection, (2) template search, sequence alignment, structure building, and refinement of each unit, (3) linker modeling, and (4) final optimization. For each target sequence, modeling units are detected by GalaxyDom[3] which runs HHsearch[4] against SCOP70[5] and PDB70. For each modeling unit, residue contacts are first predicted by CCMpred[6] from multiple sequence alignments generated by HHblits[7] on a metagenome sequence database. If the N_f value representing the effective number of related sequences[8] is larger than 20, both HHsearch and map_align[9] on PDB70 and the model database (MDB) of Baker group[9] are performed for template search. If N_f is equal to or smaller than 20, only HHsearch is run. Templates are selected by re-ranking the detected proteins using the scores of the search methods and a target difficulty score estimated by a machine learning method. Tertiary structures are built from multiple sequence alignment generated by PROMALS3D[10]. In this step, 48 models are constructed by short VTFM MD simulations with template-driven restraints and the CHARMM22 force field followed by short MD relaxations after repetitive side-chain perturbations. The models are refined using a simplified version of GalaxyRefine-CASP13 which reduces the maximum number of sampling iterations to 5 and time limit to 10 hrs from those of the original GalaxyRefine-CASP13, and five lowest-energy models are selected. If multiple modeling units are detected, orientations between the units are sampled by perturbing torsion angles of the linkers connecting the units. Final models are subject to optimization in full-atom topology to improve stereochemical properties.

Seok-refine protocol for TS targets using ProQ3D, GalaxyQA, and GalaxyRefine-CASP13

Seok-refine is a meta-server which starts with the CASP server models. All server models are first scored by ProQ3D[11], a single-model quality assessment method, and top 24 models are re-ranked by GalaxyQA, an energy-based, non-consensus model quality assessment method tested in CASP12. GalaxyQA ranks models based on an in-house knowledge-based potential called KGB[12] after local optimization with the GalaxyRefine[13,14] energy. The top model is further refined using the simplified GalaxyRefine-CASP13 mentioned in previous section with an additional modification in the structure hybridization step in which the 24 models selected by ProQ3D are used for hybridization instead of homologous protein structures. Five lowest-energy models were finally submitted.

Availability

The previous version of GalaxyTBM and GalaxyRefine are available as free web servers on the GalaxyWEB page (http://galaxy.seoklab.org). A standalone version GalaxyRefine is also downloadable (http://seoklab.github.io/GalaxyRefine).

1. Ko,J., Park,H. & Seok,C. (2012). GalaxyTBM: template-based modeling by building a reliable core and refining unreliable local regions. *BMC Bioinformatics* **13**, 198.
2. Lee,G.R., Heo,L. & Seok,C. (2018). Simultaneous refinement of inaccurate local regions and overall structure in the CASP12 protein model refinement experiment. *Proteins*. **86**, 168-176.
3. Choe,K., Heo,L., Ko,J. & Seok,C. GalaxyDom: a method to detect modeling units for protein structure prediction. *submitted.*
4. Söding,J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**, 951-960.
5. Fox,N.K., Brenner,S.E. & Chandonia,J.M. (2013). SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acid Res*. **42**, D304-D309.
6. Seemayer,S., Gruber.M. & Söding,J. (2014). CCMpred—fast and precise prediction of protein residue-residue contacts from correlated mutations. *Bioinformatics* **30**, 3128-3130.
7. Remmert,M., Biegert,A., Hauser,A. & Söding,J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature methods* **9**, 173-175.
8. Ovchinnikov,S., Kinch,L., Park,H., Liao,Y., Pei,J., Kim,D.E., Kamisetty,H., Grishin,N.V. & Baker, D. (2015). Large-scale determination of previously unsolved protein structures using evolutionary information. *Elife*, **4**, e09248.
9. Ovchinnikov,S., Park,H., Varghese,N., Huang,P.S., Pavlopoulos,G.A., Kim,D.E., Kamisetty,H., Kyrpides,N.C. & Baker, D. (2017). Protein structure determination using metagenome sequence data. *Science* **355**, 294-298.
10. Pei,J., Kim,B. & Grishin,N. (2008). PROMALS3D: a tool for multiple sequence and structure alignment. *Nucleic Acids Res*. **36**, 2295-2300.
11. Uziela,K., Menéndez Hurtado,D., Shu,N., Wallner,B. & Elofsson,A. (2017). ProQ3D: improved model quality assessments using deep learning. *Bioinformatics*, **33**, 1578-1580.
12. Heo,L. & Seok,C. A new statistical potential with consideration of solvation effects for protein simulations. *in preparation*.
13. Lee,G.R., Heo,L. & Seok,C. (2015). Effective protein model structure refinement by loop modeling and overall relaxation. *Proteins*. **84**, 293-301.
14. Heo,L., Park,H. & Seok,C. (2013). GalaxyRefine: Protein structure refinement driven by side-chain repacking. *Nucleic Acids Res.* **41**, W384-W388.

Protein structure refinement by GALAXY in CASP13

Jonghun Won, Beomchang Kang, and Chaok Seok

Department of Chemistry, Seoul National University, Seoul 08826, Republic of Korea

chaok@snu.ac.kr

In the CASP13 refinement experiment, we followed the same physics-based framework as the CASP12 protocol[1] in that several geometric operators such as anisotropic normal mode perturbation, secondary structure perturbation, and template hybridization were used to diversify sampling and final models are selected based on energy. We also added restraint energy terms derived from the initial structure with Bayesian inference.[2] In CASP13, Lorentzian-type restraints were used instead of harmonic restraints to reduce the degree of regression to the initial structure in the molecular relaxation steps. For the human predictions by Seok, oligomer interface was further considered by explicitly building oligomer environment when proper oligomer templates were available. Additional loop modeling was performed for unreliable loops detected by local quality assessment assisted by human intuition.

Methods

The overall procedure of the server protocol is similar to the CASP12 refinement server protocol.[1] First, residue-level error estimation of the initial structure was performed to detect unreliable local regions. Diverse structures were then generated by structural operators such as loop modeling, anisotropic normal mode perturbation, structure hybridization, secondary structure perturbation, and sidechain perturbation. The regions predicted to be unreliable were sampled more frequently than other regions. The generated structures were next subject to 3- or 1.2-ps molecular dynamics relaxations depending on the magnitudes of structural changes. The energy function employed for the relaxation was identical to that used in the CASP12 protocol, except that Lorentzian function is used for restraints instead of harmonic function when the initial GDT_HA is less than 60. Low-energy structures were selected and used as initial structures for the next sampling round. After iterating this procedure, five non-redundant, lowest-energy models were selected by filtering with an in-house knowledge-based potential KGB and re-ranking them with the energy function without restraints.

For human predictions, oligomer interface was considered by explicitly building oligomer structures using GalaxyHomomer[3] for targets assigned as homo-oligomer in the tertiary structure prediction experiment with detected oligomer templates. More aggressive sampling that uses no restraints during updating structures for the next round of iteration was also attempted. Unreliable loop regions were detected by local quality assessment assisted by human and were subject to loop modeling by using GalaxyLoop.[4] For target R0949, side chain atoms of H85, C159, H166, and M171 were restrained to satisfy the coordination geometry around copper.

Availability

GALAXY programs are freely available at http://galaxy.seoklab.org.

1. Lee,G.R., Heo,L. & Seok,C. (2018). Simultaneous refinement of inaccurate local regions and overall structure in the CASP12 protein model refinement experiment. *Proteins* **86**, 168-176.
2. MacCallum,J.L., Perez,A. & Dill,K.A. (2015). Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference. *Proc Natl Acad Sci U S A* **112** (22), 6985-6990.
3. Baek,M., Park,T., Heo,L., Park,C. & Seok,C. (2017). GalaxyHomomer: a web server for protein homo-oligomer structure prediction from a monomer sequence or structure. *Nucleic Acids Res* **45** (W1), W320-W324.
4. Lee,G.R., Park,H., Heo,L & Seok,C. (2014). Protein loop modeling using a new hybrid energy function and its application to modeling in inaccurate structural environments. *PLoS ONE* **9** (11), e113811.

Contact map prediction based on deep learning model driven by the topology characteristics of the query protein and parameter iterative refinement

Shi-Hao Feng[1,2], Jia-Yan Xu[1,2], Yang Yang[3], and Hong-Bin Shen[1,2]

1 - Institute of Image Processing and Pattern Recognition, ShanghaiJiaoTongUniversity, 2 - Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240, China, 3 - Department of Computer Science and Engineering, Shanghai Jiao Tong University

hbshen@sjtu.edu.cn

Accurate prediction of long-range contacts has been demonstrated to be significantly helpful in *ab initio* structure modeling since it brings strong restraints in the topology space search. Conventional methods have made great progress in contact map prediction[1-6]. Recently, a machine learning technique called deep learning is employed in this field and achieves higher precision compared to traditional methods [4,5].Although much success has been witnessed by these predictors, one typical protocol is that a general model is constructed based on a fixed training set for predicting all the query sequences. However, our preliminary tests show that the prediction accuracies of helical proteins will be much lower than the beta-strand proteins. Hence, one of our major motivations here is that the performance of the predictors should be able to be improved if the prediction model can be constructed by considering the protein topology characteristic of the predicted proteins. In CASP13, our model is based on a well-developed deep convolutional neural network called Resnet [6], which takes co-evolutional and structural features as inputs. We classify the predicted proteins into major two classes, i.e., alpha-helical proteins and beta-strand proteins, according to whether there is more than one strand segment in protein secondary structure, since the distributions of the contact maps of these two classes of proteins are very different. Then, we construct two specific models for these two classes of proteins. When predicting the contact map, we first search the homologous proteins against the newest PDB database [7] and then add these similar proteins to the corresponding training set to iterative refine the model's parameters to make them fit better to the query protein. Finally, a training process is performed and the newly obtained model is used to predict the contact map.

Methods

We classify proteins into alpha-helical proteins and beta-strand proteins, since our preliminary tests have shown that the contact distributions of these two class of proteins are significantly different, resulting in a much different performance of a single general model on these two types of proteins. The contacts in alpha-helical proteins tend to distribute sparsely while for beta-strand proteins, there are some regions where contacts are densely distributed. We construct two models for these two classes of proteins, in which the contact distribution of proteins is similar with the corresponding class of proteins. Furthermore, a model refinement steps was used in our approach, where we search the homologous proteins against the newest PDB database and use the homologous proteins to fine turn our model's parameters. These two steps are all aimed to make the contact distribution of the training samples more similar with that of the predicted proteins.

We employ the widely-used deep learning architecture Resnet as our model. The input features are composed of two parts: co-evolutional features and structural features. The co-evolutional features are PSICOV [8], EVFOLD [9], CCMPred [10], DEEPCOV [4], and PSSM [11] while the structural features are secondary structure, torsion angles, and global solvent exposure descriptors, which can be predicted by SPIDER3 [12]. All these features are combined and converted to a L×L×64 matrix as Ref 13, which will serve as the input of our model.

1. Yang J, Jin Q-Y, Zhang B, Shen H-B. R2C: improving ab initio residue contact map prediction using dynamic fusion strategy and Gaussian noise filter. Bioinformatics 2016;32(16):2435-2443.
2. Yang J, Shen H-B. MemBrain-contact 2.0: a new two-stage machine learning model for the prediction enhancement of transmembrane protein residue contacts in the full chain. Bioinformatics 2017;34(2):230-238.
3. Yin X, Yang J, Xiao F, Yang Y, Shen H-B. MemBrain: An easy-to-use online webserver for transmembrane protein structure prediction. Nano-Micro Letters 2018;10(1):2.
4. Jones DT, Kandathil SM. High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. Bioinformatics 2018;https://doi.org/10.1093/bioinformatics/bty341.
5. Wang S, Sun S, Li Z, Zhang R, Xu J. Accurate de novo prediction of protein contact map by ultra-deep learning model. PLoS computational biology 2017;13(1):e1005324.
6. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016;https://doi.org/10.1109/cvpr.2016.90.
7. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic acids research 2000;28(1):235-242.
8. Jones DT, Buchan DW, Cozzetto D, Pontil M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 2011;28(2):184-190.
9. Kaján L, Hopf TA, Kalaš M, Marks DS, Rost B. FreeContact: fast and free software for protein contact prediction from residue co-evolution. BMC bioinformatics 2014;15(1):85.
10. Seemayer S, Gruber M, Söding J. CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. Bioinformatics 2014;30(21):3128-3130.
11. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of molecular biology 1990;215(3):403-410.
12. Heffernan R, Yang Y, Paliwal K, Zhou Y. Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. Bioinformatics 2017;33(18):2842-2849.
13. Tegge AN, Wang Z, Eickholt J, Cheng J. NNcon: improved protein contact map prediction using 2D-recursive neural networks. Nucleic acids research 2009;37(suppl_2):W515-W518.

Ab initio residue contact map prediction using deep learning models

Jing Yang[1,2] and Hong-Bin Shen[1,2]

1 - Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, 2 -Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240, China

hbshen@sjtu.edu.cn

Inter-residue contacts in proteins have been widely acknowledged to be valuable for protein 3D structure prediction. Here, we present an updated version of R_2C[1] to improve the prediction of residue contacts by using deep residual neural network. In this case, we can train deeper neural network and get better learning ability as well.

Methods

The newly developed R_2C predictor consists of five ResNet-34 [2] models and the final prediction is the average of these models. The features fed into deep convolutional neural network include position-specific scoring matrix, predicted secondary structure, predicted solvent accessibility and correlated mutations. From these sequence-derived features, correlated mutation score of two residues is the most important features for enhancing the capacity of the prediction model. Concretely, we used FreeContact[3], PSICOV[4] and CCMpred[5] to detect direct couplings from multiple sequence alignment (MSA), which was generated by using HHblits[6] to search against the UniClust30 database. Experimental results demonstrate that deep learning model trained by ResNet framework performs significantly better than the initial version of shallow models.

Results

Tested on 21 CASP12 free modeling (FM) targets, the new R_2C residue contact predictor can achieve an overall accuracy of 60.8% for the top $L/5$ long-range contacts in domain level. Our R_2C web server is available at http://www.csbio.sjtu.edu.cn/bioinf/R2C/.

1. Yang,J. *et al*. (2016) R_2C: improving *ab initio* residue contact map prediction using dynamic fusion strategy and Gaussian noise filter. *Bioinfomatics*, **32**, 2435-2443.
2. He,K. *et al*. (2015) Deep residual learning for image recognition. *arXiv preprint arXiv:151203385*, 770-778.
3. Kaján,L. *et al*. (2014) FreeContact: fast and free software for protein contact prediction from residue co-evolution. *BMC Bioinfomatics*, **15**, 85.
4. Jones,D.T. *et al*. (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, **28**, 184-190.
5. Seemayer,S. *et al*. (2014) CCMpred-fast and precise prediction of protein residue-residue contacts from correlated mutations. *Bioinfomatics*, **30**, 3128-3130.
6. Remmert,M. *et al*. (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat*. *Methods*, **9**, 173-175.

Reduced Local Steric Clash and Improved Atom Packing by Segment Replacement and Constantly Changing Selective Pressures Using a Monte Carlo/Genetic Algorithm Method

David Shortle

The Johns Hopkins University School of Medicine

dshortl1@jhmi.edu

While many CASP server models have relatively little atom-atom overlap between residues separated by one or more intervening residues, all of these models retain significant to enormous amounts of atom overlap between residues i and i+1. Disregard of this energy term allows the polypeptide chain to access many new conformations which are energetically forbidden to real proteins. Disallowed changes in phi/psi/chi angles of adjacent residues provide an especially effective mechanism for generating structural changes that improve global scoring terms, since they act as new hinges or kinks. Unfortunately, subsequent conformational search enters a (much) larger space of false conformations.

Methods

Our group has spent the past 15 years developing an approach based on Monte-Carlo/Genetic Algorithm methods (MC/GA) for protein structure prediction, using the standard strategy of replacing varying length segments of a starting model with segments taken from high resolution PDB structures. When new side-chains replace the originals, atom-atom overlap is unavoidable, and surprisingly high levels typically occur. One recent modification to our protocol is to begin refinement by generating 3000-4000 very crude models by randomly joining protein segments of length 3 to 10 residues having the predicted secondary structure, low backbone atom overlap and low Ramachandran energies. All other energy terms are ignored in selecting these PDB segments as well as in picking the assembled models to be saved. After replacing the side-chains with the target sequence, these "decoys" are refined via 2 rounds of 5 generations with our MC/GA program, with constant heavy selective pressure against local and short-range atom overlap, lower Ramachandran energies, and modest but slowly increasing pressure for compactness and structural similarity over short segments to one or more realistic models.

These locally accurate but globally false decoys serve as a library of segments 1 to 30+ residues which are reassembled on to realist models (i.e., the CASP server models) acting as templates / scaffolds. Almost all server models are extensively modified in the first few generations by selecting for low local overlap plus low Ramachandran energies. A modest but increasing pressure is applied to reduce the CA-CA distance matrix error relative to the one "best" available model (initially a server model, or the CASP-provided refinement model), which is replaced later by one or more low-scoring refined models.

In three to five subsequent rounds of 4-6 generations each run on the MC/GA program, 2000-6000 new models from the previous round are used both as the library of new, lower energy segments and as the source of ensembles of 25 models, enriched for one or more parameters, to start the first generation. The expected loss of structural diversity is delayed by using a variety of tactics: (1) selection with replacement of the original model (i.e., no expansion of the ensemble with new children); (2) energy minimization by phi/psi/omega angle tweaks and small backbone bond angle changes is only employed in the last round or two; (3) one selection function drives the evolving trajectory of new conformations but a second, different survival function picks the best new model in that small set to replace the starting structure; (4) from one generation to the next the selection function is alternated between the weighted sum of 3 to 10 different composite pseudo-energy terms and the weighted sum of z-scores of a subset of these terms.

As refinement progresses, more emphasis is given to the standard global energy terms - pair potentials, solvation, hydrogen bonds - with a gradual shift from residue-specific to atom-specific terms. In parallel, increasing pressure for native-like atom packing and uniform atom density is also applied. Our experience is that conventional statistical potentials for atom-atom pair interactions or solvation do not usually improve these last two properties, which appear to be most useful for refinement proceeding at higher resolution.

Results

In summary, (1) One method, described above, was used for all predictions. (2) Extensive manual intervention was absolutely essential to achieve a synchronous drop in values of the many pseudo-energy terms throughout refinement, especially local atom overlap. Each target presented a somewhat different challenge in this regard. (3) All server models were scored for a variety of pseudo-energy terms and the secondary structure of a few server models supplemented the predictions obtained from PSIPRED. (4) A set of 40-80 server models displaying the best consensus of short-to-modest range CA-CA distances was used to initiate refinement of realistic models (i.e, round 3 as described above). (5) For the very largest CASP targets, no crude models for the first round library were generated by PDB fragment assembly as described above. Instead, the entire stage2 tarball served as the library in the first round. (5) The models submitted to CASP13 had the lowest sum of z-scores for (a) atom-atom interaction energy, (b) solvation, (c) dispersion energy, and (d, e) two packing quality terms.

Availability

All software used in this work has been written in C++ by the group leader trying to conform to best programming practices as defined in Code Complete by Steven McConnell. Since our computer programs were built around an old, proprietary object/template library purchased from RogueWave Software (Windows Version), several obstacles would have to be overcome before our code would be useable by other groups. In addition the source code needs extensive re-writing to make it understandable by other programmers.

A Novel Statistical Energy Function and Effective Conformational Search Strategy based ab initio Protein Structure Prediction

Avdesh Mishra[1], Md Tamjidul Hoque[1,*]

1 - Computer Science, University of New Orleans, 2000 Lakeshore Drive, New Orleans, LA 70148, USA

* thoque@uno.edu

In CASP13, we test our recently developed novel *ab initio* protein structure prediction (PSP) method, called 3DIGARS-PSP that utilize an effective statistical energy function, called 3DIGARS and advanced search algorithm, called KGA. The proposed method employs a memory assisted genetic algorithm (GA) derived from KGA to sample the complex energy surface of the protein folding process. The GA employs two effective operators: memory assisted crossover and mutation which are decorated with angle rotation and segment translation features to address the critical search process. Furthermore, propensities of secondary structure and dihedral angle distribution are utilized to guide the conformational search. The GA based sampling that minimizes the statistical energy function generates large-scale decoy pool. We collect top five models for each CASP13 target by clustering the ensemble of decoys and consequently submit these models to CASP13.

Methods

Protein structure in 3DIGARS-PSP are primarily represented by backbone atoms N, Cα, C and O. For each CASP13 targets, we first obtain the predicted models from I-TASSER[1]. We start by initializing some of the chromosomes of the GA population with the Cartesian coordinates of the backbone atoms of the models from I-TASSER. Next, the remaining chromosomes are initialized by single point torsion angle changes (rotation). For a guided change of the torsion angles (Φ or Ψ), we utilize the frequency of occurrence of 20 different amino acids with different Φ-Ψ angle pairs, summarized from the 4,332 high-resolution experimental structures extracted in our previous work[2]. The range of both Φ and Ψ angles for every amino acids are divided into 120 bins with an interval of 3 degrees and the frequency of the bins are updated based on the value of the Φ and Ψ angles. The frequency distribution obtained for each amino acid is further categorized into zones by looking at the cluster of the frequency values. Then, the roulette wheel selection approach is applied to select the most probable torsion angles (namely, pΦ or pΨ) belonging to the zone. Next, a random Φ or Ψ (say, rΦ or rΨ) between pΦ-3 and pΦ or pΨ and pΨ+3 is selected and rotation of the current torsion angle is performed to achieve new torsion angle, rΦ or rΨ.

In addition, the change of the torsion angles is further guided by the propensities of secondary structure (SS) types of the amino acids extracted from the 4,332 high-resolution experimental structures by running the DSSP program. The eight different SS types (E, B, H, G, I, T, S and U) given by DSSP are broadly categorized into four different SS types (H, G, and I = H; E and B = E; T and S = T; and U). The Φ-Ψ angle pair and SS types are used to obtain the index in the SS frequency table and increase the frequency count of the cell in the table by one. Later, the SS type which has the largest frequency count is assigned to the given amino acid having a certain Φ-Ψ angle. Additionally, we collect the Φ-Ψ angle pairs belonging to the H and E types and group them into helix and beta groups. We utilize the Φ-Ψ angle pairs belonging to the helix or sheet group to update the Φ or Ψ angle that results in the clash within the structure.

To generate new chromosomes (structural samples) for next generation of GA, we apply two types of conformational change operators *i)* angle rotation; and *ii)* segment translation. The mutation operation involves phi or psi angle rotation and crossover operation involves segment translation followed by phi or psi angel rotation at the crossover point. Rotation of phi and psi angles is based on an idea of rotation about an arbitrary axis. For segment translation, a set of possible crossover points are selected based on the secondary structure information. All amino acid indexes except the amino acids belonging to the beta sheet secondary structure type (either E or B) are considered as possible crossover points. This is done to preserve beta sheet regions in the structure from random changes during the crossover operation and perform more controlled changes of this region while performing mutation operation. During the crossover process, we generate four children structures from two parent structures and a structure with the best fitness saved in the memory[3].

Using the statistical energy function, decoys are generated by minimizing the potential energy using associated memory GA discussed above. Each decoy generated by 3DIGARS-PSP are then converted into the all-atom level by using Oscar-star software[4] and ranked using single-model based model quality assessment program Qprob[5], which predicts a model's quality by estimating the error of structural, physiochemical and energy-based features using probability density distributions. Next, the MUFOLD-CL[6] method is used to cluster the decoys. Then, we select the top five models in different clusters based on their Qprob rankings. The top five models are further refined using ModRefiner[7] software. Then, we use ResQ[8] method to add B-factors to the top five models before submission.

Availability

Source code, manual and example data of 3DIGARS-PSP for Linux are freely available to non-commercial use at http://cs.uno.edu/~tamjid/Software/ab_initio/v2/PSP.zip.

1. Lab,Z. I-Tasser Software, Vol. 2017, pp. http://zhanglab.ccmb.med.umich.edu/I-TASSER/.
2. Mishra,A. & Hoque,M.T. (2017). Three-Dimensional Ideal Gas Reference Sstate Based Energy Function. *Current Bioinformatics* **12**, 171-180.
3. Hoque,M.T. & Iqbal,S. (2017). Genetic algorithm-based improved sampling for protein structure prediction. *International Journal of Bio-Inspired Computation* **9**, 129-141.
4. Liang,S., Zheng,D., Zhang,C. & Standley,D.M. (2011). Fast and accurate prediction of protein side-chain conformations. *Bioinformatics* **27**, 2913-2914.
5. Cao,R. & Cheng,J. (2016). Protein single-model quality assessment by feature-based probability density functions. *Scientific Reports* **6**, 23990.
6. Zhang,J. & Xu,D. (2013). Fast algorithm for population-based protein structural model analysis. *Proteomics* **13**, 221-229.
7. Xu,D. & Zhang,Y. (2011). Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. *Biophysical Journal* **101**, 2525-2534.
8. Yang,J., Wang,Y. & Zhang,Y. (2016). ResQ: An approach to unified estimation of B-factor and residue-specific error in protein structure prediction. *Journal of Molecular Biology* **428**, 693-701.

A Novel Statistical Energy Function and Effective Conformational Search Strategy based Protein Complex Structure Prediction

Avdesh Mishra[1], Md Tamjidul Hoque[1,*]

1 - Computer Science, University of New Orleans, 2000 Lakeshore Drive, New Orleans, LA 70148, USA

* thoque@uno.edu

For the prediction of protein complex structure (or assembly prediction) in CASP13, we utilize our recently developed protein structure prediction (PSP) method, called 3DIGARS-PSP that uses an effective statistical energy function, called 3DIGARS and advanced search algorithm, called KGA. We refer to our assembly prediction method as 3DIGARS-PSP-ASSEMBLY. Our 3DIGARS-PSP method employs a memory assisted genetic algorithm (GA) extended from KGA for the conformational sampling of the protein folding process. The design of GA involves two important operators: memory assisted crossover and mutation. These operators perform the important function of angle rotation and segment translation to assist in careful sampling. Furthermore, the propensities of secondary structure and torsion angle are utilized to assist the search process. Through the memory assisted GA based sampling that minimizes the statistical energy function a large-scale ensemble of decoys are generated. Finally, the top five models for each CASP13 assembly target are collected by clustering the ensemble of decoys and consequently, these models are submitted to CASP13.

Methods

The assembly targets in CASP13 consists of more than one subunits where every subunit has a corresponding fasta sequence. For each assembly targets in CASP13, we prepare a single fasta sequence by combining the fasta sequences of the subunits by adding 20 Glycine (GLY or G) amino acids in between the fasta sequences. Glycine amino acid is used to combine the fasta sequences of the subunits because of its smallest size of the side chain among 20 standard amino acids. The combined fasta sequence is then used to obtain the predicted models from I-TASSER[1]. The prediction of the 3D structure of assembly target starts by initializing some of the chromosomes of the GA population with the Cartesian coordinates of the backbone atoms of the models obtained from I-TASSER. The rest of the chromosomes are filled by single point torsion angle changes (rotation). For the informed change of the torsion angles (Φ or Ψ), the occurrence frequency of 20 standard amino acids with different Φ-Ψ angle pairs are constructed from the 4,332 high-resolution experimental structures extracted in our previous work[2]. To obtain the frequency of distribution of 20 standard amino acids, the Φ and Ψ angle range is divided into 120 bins with an interval of 3 degrees and the frequency of the bins are updated based on the value of the Φ and Ψ angles of every amino acid in the protein. The frequency distributions are further categorized into zones by looking at the cluster of the frequency values. Consequently, using the roulette wheel selection method the most probable torsion angle (namely, pΦ or pΨ) of the zone is extracted and a random angle around this angle is selected as a new torsion angle.

Moreover, the propensities of secondary structure (SS) types of the amino acids is also extracted from the same experimental structures used above by running the DSSP program to guide the torsion angle rotation. The SS types given by DSSP are broadly categorized into four different SS types (H, G, and I = H; E and B = E; T and S = T; and U). The torsion angle pair and SS types of the amino acids in protein are used to obtain the SS distribution. Later, this distribution of SS is used such that the SS type which has the largest frequency count is assigned to the given amino acid having the certain Φ-Ψ angle. Furthermore, the Φ-Ψ angle pairs corresponding to the H and E types are grouped into helix and beta groups and are consequently used to update the Φ or Ψ angle that results in a clash within the structure.

The chromosomes (models) for the next generation of GA are obtained by two different types of structural change operators: *i)* angle rotation, and *ii)* segment translation. The mutation in GA involves torsion angle rotation and crossover involves segment translation followed by torsion angel rotation at the crossover point. Torsion angle rotation technique is based on the principle of rotation about an arbitrary axis. On the other hand, crossover in GA performs segment translation where all the amino acid indexes that are not SS type E or B are considered as possible crossover points. This is done to avoid random changes in the beta sheet region and make more appropriately guided change during the mutation operation. The children structures in the crossover process are generated from two parent structures and a structure with the best fitness saved in the memory[3].

The decoys generated by the conformational change through memory assisted GA guided by the statistical energy function are then converted into the all-atom level by using Oscar-star software[4]. The large-scale pool of decoys are clustered into five different cluster groups, at least 5Å apart among each other based on the average root-mean-square deviation (RMSD). Then, we select the top five models in different clusters based on the 3DIGARS energy score ranking. The subunits of the top five models are further refined using the ModRefiner[5] software. Then, we use the ResQ[6] method to add B-factors to the subunits of the top five models. Finally, the models of the subunits are combined together in CASP13 assembly format before submission.

Availability

Source code, manual and example data of 3DIGARS-PSP for Linux are freely available, for non-commercial use, at http://cs.uno.edu/~tamjid/Software/ab_initio/v2/PSP.zip.

1. Lab, Z. I-Tasser Software, Vol. 2017, pp. http://zhanglab.ccmb.med.umich.edu/I-TASSER/.
2. Mishra, A. & Hoque, M. T. (2017). Three-Dimensional Ideal Gas Reference State Based Energy Function. *Current Bioinformatics* **12**, 171-180.
3. Hoque, M. T. & Iqbal, S. (2017). Genetic algorithm-based improved sampling for protein structure prediction. *International Journal of Bio-Inspired Computation* **9**, 129-141.
4. Liang, S., Zheng, D., Zhang, C. & Standley, D. M. (2011). Fast and accurate prediction of protein side-chain conformations. *Bioinformatics* **27**, 2913-2914.
5. Xu, D. & Zhang, Y. (2011). Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. *Biophysical Journal* **101**, 2525-2534.
6. Yang, J., Wang, Y. & Zhang, Y. (2016). ResQ: An approach to unified estimation of B-factor and residue-specific error in protein structure prediction. *Journal of Molecular Biology* **428**, 693-701.

A Novel Statistical Energy Function and Effective Conformational Search Strategy based Refinement of Protein Structure

Avdesh Mishra[1], Md Tamjidul Hoque[1,*]

1 - Computer Science, University of New Orleans, 2000 Lakeshore Drive, New Orleans, LA 70148, USA

* thoque@uno.edu

In CASP13, we test our recently developed protein structure prediction (PSP) method, called 3DIGARS-PSP that utilize an effective statistical energy function, called 3DIGARS and advanced search algorithm, called KGA for the refinement of protein structure. We refer to our refinement method as 3DIGARS-PSP-REFINE. It employs a memory assisted genetic algorithm (GA) derived from KGA to sample the energy hypersurface of the protein folding process. The GA deploys two operators: memory assisted crossover and mutation which perform angle rotation and segment translation to address the critical search process. Furthermore, secondary structure and dihedral angle propensities are utilized to guide the search process. A large-scale ensemble of decoys is generated by the GA based sampling that minimizes the statistical energy function. We collect top five models for each CASP13 refinement target by clustering the ensemble of decoys and consequently submit these models to CASP13.

Methods

For each refinement targets, we first obtain the partial initial structure provided by the CASP13. Next, we predict the structure for complete fasta sequence of refinement targets provided by the CASP13 using 3DIGARS-PSP and obtain top five models. Then, the partial initial structure is merged with the predicted complete models using segment translation technique. The segment translation is performed in order to preserve the original orientation of the initial structure provided by the CASP13. Later, the merged models are further refined using the ModRefiner[7] software to ensure that the models are free from steric clashes. Ultimately, the refined five models are used as initial seed in 3DIGARS-PSP to perform refinement.

The Cartesian coordinates of the backbone atoms of the initial seeds are used to initialize five chromosomes of the GA population. The remaining chromosomes are filled by single point torsion angle changes (rotation). For a guided change of the torsion angles (Φ or Ψ), the occurrence frequency of 20 different amino acids with different Φ-Ψ angle pairs are summarized from the 4,332 high-resolution experimental structures extracted in our previous work[2]. The Φ and Ψ angle range of 20 standard amino acids are divided into 120 bins with an interval of 3 degrees and the frequency of the bins are updated based on the value of the Φ and Ψ angles of every amino acid in the protein. The frequency distribution of the amino acids is further categorized into zones by looking at the cluster of the frequency values. The most probable torsion angle (namely, pΦ or pΨ) of the zone is extracted using the roulette wheel selection method and a random angle around this angle is selected as a new torsion angle.

Likewise, the propensities of secondary structure (SS) types of the amino acids is also extracted from the same dataset mentioned above by running the DSSP program to guide the torsion angle change. Eight different SS types given by DSSP are broadly categorized into four different SS types (H, G, and I = H; E and B = E; T and S = T; and U). The Φ-Ψ angle pair and SS types are used to obtain the SS distribution. Later, the SS type which has the largest frequency count is assigned to the given amino acid having the certain Φ-Ψ angle. Additionally, the Φ-Ψ angle pairs belonging to the H and E types are grouped into helix and beta groups. These Φ-Ψ angle pairs of helix or sheet group are later used to update the Φ or Ψ angle that results in a clash within the structure.

To generate chromosomes (models) for the next generation of GA, two different types of conformational change operators are used *i)* angle rotation; and *ii)* segment translation. The mutation operation involves torsion angle rotation and crossover operation involves segment translation followed by torsion angel rotation at the crossover point. Torsion angle rotation is based on the principle of rotation about an arbitrary axis. On the other hand, for segment translation, all the amino acid indexes that are not SS type E or B are considered as possible crossover points. This is done to preserve beta sheet regions in the structure from random changes during the crossover operation. The children structures in the crossover process are generated from two parent structures and a structure with the best fitness saved in the memory[3].

The decoys are generated by minimizing the potential energy using associated memory GA discussed above and the statistical energy function, called 3DIGAS. Each decoy generated are then converted into the all-atom level by using Oscar-star software[4] and ranked using single-model based model quality assessment program Qprob[5]. Next, the MUFOLD-CL[6] method is used to cluster the decoys. Then, we select the top five models in different clusters based on their Qprob rankings. The top five models are further refined using the ModRefiner[7] software. Then, we use the ResQ[8] method to add B-factors to the top five models. Finally, partial models of top five models containing only residues that are required by CASP13 are created before submission.

Availability

Source code, manual and example data of 3DIGARS-PSP for Linux are freely available, for non-commercial use, at http://cs.uno.edu/~tamjid/Software/ab_initio/v2/PSP.zip.

1. Lab,Z. I-Tasser Software, Vol. 2017, pp. http://zhanglab.ccmb.med.umich.edu/I-TASSER/.
2. Mishra,A. & Hoque,M.T. (2017). Three-Dimensional Ideal Gas Reference State Based Energy Function. *Current Bioinformatics* **12**, 171-180.
3. Hoque,M.T. & Iqbal,S. (2017). Genetic algorithm-based improved sampling for protein structure prediction. *International Journal of Bio-Inspired Computation* **9**, 129-141.
4. Liang,S., Zheng,D., Zhang,C. & Standley,D.M. (2011). Fast and accurate prediction of protein side-chain conformations. *Bioinformatics* **27**, 2913-2914.
5. Cao,R. & Cheng,J. (2016). Protein single-model quality assessment by feature-based probability density functions. *Scientific Reports* **6**, 23990.
6. Zhang,J. & Xu,D. (2013). Fast algorithm for population-based protein structural model analysis. *Proteomics* **13**, 221-229.
7. Xu,D. & Zhang,Y. (2011). Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. *Biophysical Journal* **101**, 2525-2534.
8. Yang,J., Wang,Y. & Zhang,Y. (2016). ResQ: An approach to unified estimation of B-factor and residue-specific error in protein structure prediction. *Journal of Molecular Biology* **428**, 693-701.

Contact map prediction by deep residual fully convolutional neural network with only evolutionary coupling features derived from deep multiple sequence alignment

Chengxin Zhang[1], Yang Li[1,2], Dongjun Yu[2], Yang Zhang[1]

1 Department of Computational Medicine and Bioinformatics, Department of Biological Chemistry, University of Michigan, 100 Washtenaw Ave, Ann Arbor, MI 48109, 2 School of Computer Science and Engineering, Nanjing University of Science and Technology, Xiaolingwei 200, Nanjing, China, 210094

yangzhanglab@umich.edu

TripletRes was developed to take full advantage of evolutionary coupling features and deep convolutional neural networks, both of which are becoming important components of contact map predictors. A deep multiple sequence alignment (MSA) building pipeline was also developed to extract more discriminative evolutionary coupling features.

Methods

Multiple sequence alignment is the fundamental element to generate evolutionary couplings. In this pipeline, a high-quality multiple sequence alignment was obtained by a hierarchical sequence searching protocol. The query sequence was firstly searched against UniClust30[1] database by HHblits[2], followed by jackhmmer[3] searching through Uniref90. HHblits was again used to search the sequences obtained by jackmmer. The concatenated MSA was further enriched by hmmsearch[3] through MetaClust[4] database.

We derived three features from the deep MSA. The first feature is the ridge estimation of inverse covariance matrices from one-hot-encoded MSA. The second is the coupling parameter matrix of pseudo-likelihood maximization[5]. Covariance matrix is the last feature considering that marginal relationships may also help. Each feature is a matrix with the size of (21*L) by (21*L) for a protein sequence with L amino acids.

For each feature, the entries of the 21 by 21 sub-matrix of a corresponding amino acid pair are the descriptors and were fed into a convolutional transformer conducted by a fully convolutional neural network with residual architecture[6]. The transformed features of three feature inputs are concatenated together as the input of another deep residual fully convolutional neural network that outputs the predicted contact map. The three neural networks transforming the input features and the last neural network that outputs the final prediction are trained together end-to-end. This is TripletRes' main difference from the ResTriplet server, which trains the first three neural networks separately and stacked them together by another neural network.

ThreaDom[7] was employed for domain boundary prediction. For a multi-domain sequence, contact map for full length is firstly predicted and the individual intra-domain contacts are replaced with predicted contact maps based on their own domain sequences from the same TripletRes predictor, without renormalizing the confidence scores.

Availability

The web server of TripletRes is available at https://zhanglab.ccmb.med.umich.edu/TripletRes.

1. Mirdita M, von den Driesch L, Galiez C, Martin MJ, Söding J, Steinegger M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic acids research* 2016;**45**(D1):D170-D176.
2. Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature methods* 2012;**9**(2):173.
3. Eddy SR. Accelerated profile HMM searches. PLoS computational biology 2011;7(10):e1002195.

4. Steinegger M, Söding J. Clustering huge protein sequence sets in linear time. *Nature communications* 2018;**9**(1):2542.
5. Seemayer S, Gruber M, Söding J. CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. *Bioinformatics* 2014;**30**(21):3128-3130.
6. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016. p 770-778.
7. Xue Z, Xu D, Wang Y, Zhang Y. ThreaDom: extracting protein domain boundary information from multiple threading alignments. *Bioinformatics* 2013;**29**(13):i247-i256.

Use of improved UNRES force field and replica-exchange molecular dynamics in physics-based template-free and data-assisted prediction of protein structures

E.A. Lubecka[1,2], A.G. Lipska[2], A.K. Sieradzan[2], K. Zięba[2], A.S. Karczyńska[2], C. Sikorska[2], U. Uciechowska[2], S.A. Samsonov[2], P. Krupa[3], M.A. Mozolewska[4], Ł. Golon[2], A. Giełdoń[2], C. Czaplewski[2], R. Ślusarz[2], M. Ślusarz[2], and A. Liwo[2*]

1 - Institute of Informatics, Faculty of Mathematics, Physics, and Informatics, University of Gdańsk, Wita Stwosza 57, 80-308 Gdańsk, Poland, 2 - Faculty of Chemistry, University of Gdańsk, Wita Stwosza 63, 80-308 Gdańsk, Poland, 3 - Institute of Physics, Polish Academy of Sciences, Aleja Lotników 32/46, Warsaw, PL-02668, Poland, 4 - Institute of Computer Science, Polish Academy of Sciences, ul. Jana Kazimierza 5, Warsaw 01-248, Poland

adam.liwo@ug.edu.pl

Physics-based approaches are, so far, less efficient than knowledge-based approaches in the prediction of protein structures; however, their advantage is independence of structural databases. In the physics-based approaches**,** the predicted structure is sought as one with the lowest free energy at physiological conditions. Use of all-atom approaches still involves prohibitively high simulation cost which, however, can be largely reduced when coarse-grained protein models are used.

In the last several years, we have been developing the physics-based united-residue (UNRES) force field for physics-based prediction of protein structures and large-scale simulations of protein folding, together with a variety of methods for searching the conformational space[1]. Recently we introduced various improvements in UNRES. This new force field has been tested in the present CASP experiment.

Methods

In the UNRES model[1], a polypeptide chain is represented by a sequence of alpha-carbon atoms connected by virtual bonds with attached side chains. Two interaction sites are assigned to each amino-acid residue: the united peptide group (p) located in the middle of two consecutive alpha-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 the last year, we developed new functional expressions for the backbone virtual-bond-torsional and correlation potentials, in which the dependence on the backbone virtual-bond angles adjacent to a given virtual-bond-dihedral angle is introduced, this resulting in major improvement of the calculated β- and loop structures[2]. The force field was subsequently calibrated with 9 proteins of different secondary structure and size from 20 to 70 residues, by using the maximum-likelihood method developed in our laboratory[3].

The structures of the target proteins were predicted by the following four-stage procedure. First, UNRES was employed to carry out Multiplexed Replica Exchange Molecular Dynamics (MREMD)[4] for target proteins. To speed up the search for larger proteins, weak restraints were imposed on secondary structure based on secondary structure prediction by PSIPRED[5]; raw PSIPRED data were used and, consequently, the respective restraint function was bimodal with one minimum in the alpha-helical and another one in the extended region, well-depths depending on PSIPRED-determined probabilities. Second, based on MREMD simulation results, Weighted-Histogram Analysis Method (WHAM) was used to calculate the relative free energy of each structure of the last section of MREMD simulation[1]. Third, cluster analysis was employed to cluster the structures from an MREMD simulation. Five clusters with the lowest free energies were chosen as prediction candidates. Finally, in the fourth stage, the conformations closest to the respective average structures corresponding to the found clusters were converted to all-atom structures using the PULCHRA[6] and SCWRL[7] algorithms. Subsequently, the AMBER14 package[8] with the ff14SB force field and GBSA implicit-solvent model was used to refine the all-atom models, by carrying out 500 minimization steps followed by 0.3 ps of molecular dynamics, with restraints on the secondary structure and positional restraints from the parent UNRES structure. Such refined all-atom structures were submitted to the CASP website.

All types of 3D-structure predictions were run (regular, data-assisted, and refinement). In data-assisted predictions (using SAXS/SANS, cross-linking, and simulated and real NMR data), appropriate penalty terms were added to the target functions. A special penalty function was engineered to handle ambiguous NMR restraints.

Results

We postpone the assessment of the approach until the official release of CASP13 results.

Availability

The UNRES package is available at 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, Chapter 8, pp. 107-122.
2. Sieradzan,A.K., Makowski,M., Augustynowicz,A. & Liwo, A. (2017) A general method for the derivation of the functional forms of the effective energy terms in coarse-grained energy functions of polymers. I. Backbone potentials of coarse-grained polypeptide chains. *J. Chem. Phys.*, **146**, 124106.
3. Krupa,P., Hałabis,A., Żmudzińska,W., Ołdziej,S., Scheraga,H.A., Liwo,A. (2017) Maximum likelihood calibration of the UNRES force field for simulation of protein structure and dynamics. *J. Chem. Inf. Model*. **57**, 2364-2377.
4. 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. *J Chem. Theory Comput.* **5**, 627-640.
5. McGuffin,L.J., Bryson,K.& Jones,D.T. (2000) The PSIPRED protein structure prediction server. *Bioinformatics*, **16**, 404-405.
6. Rotkiewicz,P. & Skolnick,J. (2008) Fast procedure for reconstruction of full-atom protein models from reduced representations. *J. Comput. Chem.*, **29**, 1460-1465.
7. Wang,Q., Canutescu,A.A. & Dunbrack,R.L. (2008) SCWRL and MolIDE: Computer programs for side-chain conformation prediction and homology modeling. *Nat. Protoc.* **3**,1832-1847.
8. Case,D.A. et al. (2014), AMBER 14, University of California, San Francisco.

Protein model quality estimation by clustering-based method SARTclust in CASP13

Kun-Sop Han and Myong-Ho Choe

Department of Life Science, University of Science, Unjong-District, Pyongyang, DPR Korea

hks1981@star-co.net.kp

Quality Assessment (QA) of protein models is an essential component in any protein structure prediction method and is important for determining its usefulness for specific application. We participated in QA category of CASP13 with two prediction methods. The method described here is labeled as group "UOSHAN", number 194. This method is a new clustering-based quality assessment program that predicts local as well as global quality of protein models. For global and local score, all server models submitted from a target protein are ranked according to their SART_G scores, and a reference set composed of n top-ranked models is formed. A given model to be assessed is compared with all models of the reference set using TMscore[2]. For global score, n GDT_TS scores produced from comparison are focused on. The clustering-based global quality score, SARTclust_G, is the SART_G-weighted mean of n GDT_TS scores. For local score, n Cα distances (d) between the corresponding residues, are computed. The distance, d, is converted to the S-score with distance threshold d_0=3.8Å, S =1/(1+$(d/d_0)^2$). Next, the SART_G-weighted mean (S_Weight) of n S-scores is calculated. The per residue distance deviation, SARTclust_L = d_0 (1/ S_Weight - 1)$^{1/2}$. We put all SARTclust_L >15Å to 15Å.

Methods

1. SARTclust_G: Clustering-based global quality score

We develop a new clustering-based global quality score SARTclust_G for global quality estimation of protein model. The procedure of calculating SARTclust_G is given below.

Step 1. All server models submitted from a target protein are ranked according to their SART_G scores. A reference set U composed of n top-ranked models is formed. In case of CASP13 stage1, n=11, and for stage2, n = 21.

Step 2. A given model, i, (model to be assessed) is compared with n models of reference set U using TMscore, resulting in n $GDT\_TS_{i,\,u}$ scores.

Step 3. The clustering-based global quality score of the model, i, $SARTclust\_G_i$, is SART_G-weighted mean of n $GDT\_TS_{i,u}$ scores.

$$\text{SARTclust\_G}_i = \frac{\sum_{u=1}^{n}(GDT\_TS_{i,u} \times SART\_G_u)}{\sum_{u=1}^{n} SART\_G_u}, \quad \text{if } GDT\_TS_{i,u} \neq 100,\ u=1\sim n. \quad (1)$$

2. SARTclust_L: Clustering-based local quality score

We develop a new clustering-based local quality score SARTclust_L for local quality estimation of protein model. The procedure of calculating the SARTclust_L is described below.

Step 1. It is the same as Step 1 in SARTclust_G calculation

Step 2. A given model (model to be assessed), i, is compared with n models of the reference set U using TMscore. The Cα distance, $d_{i,\,u,\,t}$, between the corresponding residue, t, is computed after superposition of the given model, i, and the model, u, in the reference set U.

Step 3. The distance, $d_{i,\,u,\,t}$, is converted to the S-score with distance threshold d_0=3.8Å, $S_{i,\,u,\,t}$ =1/(1+$(d_{i,\,u,\,t}/d_0)^2$). Each residue in the given model has n $S_{i,\,u,\,t}$ from superposition between the given model and n reference models.

Step 4. The SART_G-weighted mean of n $S_{i,\,u,\,t}$ scores is calculated.

$$\text{S\_Weight}_{i,\,t} = \frac{\sum_{u=1}^{n}(S_{i,u,t} \times SART\_G_u)}{\sum_{u=1}^{n} SART\_G_u}, \qquad \text{if } d_{i,\,u,\,t} \neq 0,\ u=1\sim n. \qquad (2)$$

Step 5. The per residue distance deviation, SARTclust_L$_{i,\,t}$, is calculated by SARTclust_L$_{i,\,t}$ = d_0 (1/ S_Weight$_{i,\,t}$ -1)$^{1/2}$. We put all SARTclust_L$_{i,\,t}$>15Å to 15Å.

Results

We use 3 metrics (i.e. average per target Pearson correlation coefficient, average per target quality loss and AUC (cutoff = 0.5 GDT_TS)) to assess the global prediction methods for 91 CASP11 targets.

Table 1. Comparison of SARTclust_G with clustering-based global QA methods in CASP11

Methods	stage1			stage2		
	Average Pearson	Average quality loss	AUC	Average Pearson	Average quality loss	AUC
410(Pcons-net)	0.799	0.026	0.982	0.625	0.052	0.981
SARTclust_G	0.796	0.052	0.985	0.696	0.051	0.986
268(MULTICOM-REFINE)	0.791	0.057	0.977	0.543	0.076	0.983
171(MUFOLD-SERVER)	0.781	0.056	0.982	0.555	0.074	0.567
038(nns)	0.780	0.032	0.970	0.510	0.080	0.952
347(Wallner)	0.754	0.056	0.983	0.616	0.048	0.980
239(ModFOLDclust2)	0.713	0.056	0.980	0.537	0.072	0.983

We use 2 metrics (i.e. average per model Pearson correlation coefficient and MCC (cutoff = 3.8Å)) to assess the predicted local quality scores against the real deviations for 63 CASP11 targets.

Table 2. Comparison of SARTclust_L with clustering-based local QA methods in CASP11

Methods	stage1		stage2	
	Average Pearson	MCC	Average Pearson	MCC
410(Pcons-net)	0.516	0.706	0.631	0.715
347(Wallner)	0.494	0.703	0.635	0.730
SARTclust_L	0.437	0.721	0.635	0.761
239(ModFOLDclust2)	0.428	0.684	0.627	0.753
268(MULTICOM-REFINE)	0.004	0.116	0.026	0.103

Availability

Manuscript for SARTclust is in preparation.

1. Kryshtafovych,A., Barbato,A., Fidelis,K., Monastyrskyy,B., Schwede,T., and Tramontano,A. (2014) Assessment of the assessment: Evaluation of the model quality estimates in CASP10. *Proteins*, **82**, 112-126.
2. Zhang,Y., Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**(4), 702-710.

Modeling CAPRI Targets 137 – 159 by Template-Based and Free Docking

P.J. Kundrotas, S. Belkin, D. Chakravarty, V.D. Badal and I.A. Vakser

Computational Biology Program and Department of Molecular Biosciences, University of Kansas, Lawrence, KS

vakser@ku.edu and pkundro@ku.edu

Experimentally determined structures account only for a fraction of known proteins. Thus, protein docking has to rely primarily on modeled structures of the individual proteins, which are typically less accurate than the experimentally determined ones.[1] The CASP-CAPRI rounds provide a unique opportunity to test docking on model structures generated by the CASP participants.

Methods. We adopted docking protocols successfully used in the previous CASP12-CAPRI37 round.[2] We used all CASP Stage 2 models of the individual proteins, except those with the loose packing. For the template-based (TB) docking, the structural alignment of the proteins was performed by TM-align.[3] The TB docking predictions were scored by a combination of structure similarity scores, normalized AACE18 values for the interface, fraction of shared target/template contacts, target/template interface sequence identity, interface solvation score, and the extent of clashes in the unrefined predictions.[2] The template free (TF) docking was by GRAMM[4] scored by AACE18 potential.[5] To constrain the docking, an automated procedure was used to mine literature for the binding site residues.[6] The final predictions were minimized by TINKER.[7]

Results. At the time of the abstract submission the assessment results were not available. Thus, we present only the modeling protocols (the summary is in Table 1). For T149 (two 5-domain 1589-residue chains), the challenge was to predict also the interdomain arrangement. For this target, additional experimental data (SAXS and cross-linking mass-spectrometry) was provided (T150 and T151). Another big target, T159, consisted of 18 subunits of three proteins. For each *n*-homomeric target (except T149), the procedure performed spatial rearrangement of the target protein to match the monomers of the co-crystallized templates. The templates were either from the full-structure library[8] or from a smaller library generated for a particular target on demand. The on-demand templates were (a) identified by HHpred[9,10] as likely (> 90% probability) templates for the target monomer, and (b) had oligomeric state in the biounit matching that of the target. For T149, we chose CASP and TF docking models with the best correspondence to the order of domains within and between the chains described in the literature.[11] For assisted targets T150 and T151, we manually adjusted models of T149, to satisfy the SAXS and cross-linking constraints. For heterodimers (T142, T155-157), we looked for common HHpred templates, when templates for the target monomers were identified either as interacting chains within a PDB entry, or non-overlapping parts of the same chain (T142). If no reliable HHpred templates were found, we performed TB docking using the large template library, followed by TF docking. For these targets, we performed cross-docking of all selected CASP Stage 2 models. For hetero-tetramer T146, we constructed TB models of the heterodimer based on a monomer of template 1g29 using various CASP models. We performed TB docking of the resulting heterodimers, based on homodimeric 1g29. For the big 18-mer T159, we identified several templates that form stacked 6-protein rings and performed TB docking of various CASP models using these templates.

Availability. The docking procedures and libraries used in this round are partially available from the DOCKGROUND resource at http://dockground.compbio.ku.edu.

Table 1. Docking summary

CAPRI/CASP target	Assembly	Experimental method	Number of residues	Number of HHpred templates	Number of docking models		
					TB	TF	Manual
T137/T0965	A2	X-ray	334	29	12	-	-
T138/T0966	A2	X-ray	494	2	2	8	-
T139/T0961	A4	X-ray	505	45	22	-	-
T140/T0973	A2	X-ray	146	106	11	-	-
T141/T0976	A2	X-ray	252	34	18	4	1
T142/H0974	AB	X-ray	72/95	49	21	-	-
T143/T0983	A2	X-ray	245	23	10	-	-
T144/T0984	A2	EM	752	9	27	-	-
T146/H0993	A2B2	X-ray	269/109	1	26	-	-
T147/T0995	A8	EM	330	28	19	-	-
T148/T0997	A2	X-ray	228	6	5	11	-
T149/T0999	A2	X-ray	1589	-	-	6	7
T152/T1003	A2	X-ray	474	33	16	-	-
T153/T1006	A2	X-ray	79	29	25	-	-
T154/T1009	A2	X-ray	718	27	28	-	-
T155/H1015	AB	X-ray	89/129	6	15	-	-
T156/H1017	AB	X-ray	111/129	-	15	16	-
T157/H1019	AB	X-ray	58/88	2	23	8	-
T158/H1020	A3	EM	577	26	26	-	-
T159/H1021	A6B6C6	EM	149/354/295	2	10	-	-

1. Anishchenko I, Kundrotas PJ, Vakser IA. Modeling complexes of modeled proteins. Proteins. 2017;85:470-8.
2. Kundrotas PJ, Anishchenko I, Badal VD, Das M, Dauzhenka T, Vakser IA. Modeling CAPRI targets 110-120 by template-based and free docking using contact potential and combined scoring function. *Proteins*. 2018;**86** Suppl 1:302-10.
3. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* 2005;**33**:2302-9.
4. Vakser IA. Protein docking for low-resolution structures. *Protein Eng.* 1995;**8**:371-7.
5. Anishchenko I, Kundrotas PJ, Vakser IA. Contact Potential for Structure Prediction of Proteins and Protein Complexes from Potts Model. *Biophys J*. 2018;**115**:809-21.
6. Badal VD, Kundrotas PJ, Vakser IA. Text Mining for Protein Docking. *PLoS Comput Biol.* 2015; **11**:e1004630.
7. Ren P, Wu C, Ponder JW. Polarizable Atomic Multipole-based Molecular Mechanics for Organic Molecules. *J Chem Theory Comput.* 2011;**7**:3143-61.
8. Anishchenko I, Kundrotas PJ, Tuzikov AV, Vakser IA. Structural templates for comparative protein docking. *Proteins.* 2015;**83**:1563-70.
9. Remmert M, Biegert A, Hauser A, Soding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods.* 2011;**9**:173-5.
10. Soding J. Protein homology detection by HMM-HMM comparison. Bioinformatics. 2005;21:951-60.
11. Hawkins AR, Smith M. Domain structure and interaction within the pentafunctional arom polypeptide. *Eur J Biochem*. 1991;**196**:717-24.

Modeling of Protein Complexes in CASP13 and CAPRI Round 46

J. Dapkūnas, K. Olechnovič and Č. Venclovas

Institute of Biotechnology, Life Sciences Center, Vilnius University

justas.dapkunas@bti.vu.lt, venclovas@ibt.lt

To model multimeric protein structures, we used a number of protein structure prediction tools, but most extensively relied on two methods, developed in our laboratory: the PPI3D web server[1] for searching of multimeric templates and VoroMQA[2] for accuracy estimation, scoring and ranking of structural models.

Methods

Structures of targets, for which multimeric templates were available, were predicted using the same comparative modeling workflow as in previous joint CASP12-CAPRI experiment[3]. The initial search for templates was performed by the PPI3D web server using BLAST and PSI-BLAST[1;4], followed by HHpred implemented in the MPI Bioinformatics Toolkit[5]. If any templates were found, initial multimeric structural models were generated using MODELLER[6]. These initial models were then refined by fragment-guided molecular dynamics[7] or by template-based docking utilizing the best CASP server models as monomers.

If templates for protein complexes could not be identified using the above described procedures, the following two modeling strategies were attempted. First, free docking of top 5 selected monomeric CASP server models was done using Hex[8] or PyDockWEB[9] for hetero-complexes and Sam[10] for homomultimers. Alternatively, in the cases of bacterial contact-dependent inhibition toxin-antitoxin targets, all structures of toxin-antitoxin complexes were downloaded from the Protein Data Bank, and CASP server models were aligned on the corresponding subunits by TM-align[11]. The resulting structural alignments were visually inspected to select promising templates for comparative modeling. Afterwards all generated models were ranked and clustered, and the model with the highest rank was selected from each cluster.

For model evaluation and selection we employed VoroMQA ("Voronoi tessellation-based Model Quality Assessment")[2]. This method combines the idea of knowledge-based statistical potentials with the use of the Voronoi tessellation of atomic balls. VoroMQA uses contact areas for describing and integrating both explicit interactions between protein atoms and implicit interactions of protein atoms with the solvent. VoroMQA produces scores at atomic, residue and global levels, thus it can also calculate two types of interface-specific scores: (1) quality score derived from the local scores of the atoms that participate in inter-chain contacts, and (2) the estimate of raw pseudo-energy that is a sum of products of inter-chain contact areas and the corresponding values of the knowledge-based statistical potential. We used both the whole-structure and the interface-specific VoroMQA scores to perform a tournament-based ranking of candidate models[3]. The same ranking procedure was also employed for selecting best models from the sets provided for the CAPRI scorers. These sets contained a considerable fraction of models with high number of clashes. Such models were filtered out before applying VoroMQA.

Results

Structural templates were identified for 20 of 43 multimeric CASP13 targets (12 of 20 CAPRI targets), and models for them were generated and refined using the standard procedure. Partial templates were also available for other targets. For example, some heteromeric targets had templates only for some of the interfaces, therefore the whole complex was assembled using docking (H0953, H0993/CAPRI T146), or by structurally aligning parts of the complex on a low resolution template structure (H1021/CAPRI T159, H1022). Large proteins usually had structural templates only for inter-domain homomers (T0960, T0963, T0981, T0999/CAPRI T149, T1000, T1004, T1020/CAPRI T158). In most of the latter cases we encountered difficulties in modeling inter-domain linkers and orientation between domains of individual subunits.

Availability

The PPI3D web server is available at http://bioinformatics.ibt.lt/ppi3d/. The VoroMQA web application is available at http://bioinformatics.ibt.lt/wtsam/voromqa. Standalone VoroMQA software for Linux is included in the Voronota package available from http://bitbucket.org/kliment/voronota/downloads.

1. Dapkūnas,J., Timinskas,A., Olechnovič,K., Margelevičius,M., Dičiūnas,R. & Venclovas,Č. (2017). The PPI3D web server for searching, analyzing and modeling protein-protein interactions in the context of 3D structures. *Bioinformatics* **33,** 935–937.
2. Olechnovič,K. & Venclovas,Č. (2017). VoroMQA: Assessment of protein structure quality using interatomic contact areas. *Proteins* **85,** 1131–1145.
3. Dapkūnas,J., Olechnovič,K. & Venclovas,Č. (2018). Modeling of protein complexes in CAPRI Round 37 using template-based approach combined with model selection. *Proteins* **86 Suppl 1,** 292–301.
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. *Nucleic Acids Res.* **25,** 3389–3402.
5. Zimmermann,L., Stephens, A., Nam,S.-Z., Rau,D., Kübler,J., Lozajic,M., Gabler,F., Söding,J., Lupas,A.N. & Alva,V. (2018). A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *J. Mol. Biol.* **430,** 2237–2243.
6. Šali,A. & Blundell,T.L. (1993). Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J. Mol. Biol.* **234,** 779–815.
7. Zhang,J., Liang,Y. & Zhang,Y. (2011). Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling. *Structure* **19,** 1784–1795.
8. Ritchie,D.W. & Kemp,G.J. (2000). Protein docking using spherical polar Fourier correlations. *Proteins 39*, 178–194.
9. Jiménez-García,B., Pons,C. & Fernández-Recio,J. (2013). pyDockWEB: a web server for rigid-body protein-protein docking using electrostatics and desolvation scoring. *Bioinformatics* **29,** 1698–1699.
10. Ritchie,D.W. & Grudinin,S. (2016). Spherical polar Fourier assembly of protein complexes with arbitrary point group symmetry. *J. Appl. Cryst.* **49,** 158–167.
11. Zhang,Y. & Skolnick,J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33,** 2302–2309.

Model Quality Assessment and Selection Using VoroMQA

K. Olechnovič and Č. Venclovas

Institute of Biotechnology, Life Sciences Center, Vilnius University

kliment.olechnovic@bti.vu.lt

We participated in CASP13 with two automated model accuracy estimation servers, VoroMQA-A and VoroMQA-B, and a model selection method, VoroMQA-select, registered as a regular tertiary structure prediction group. All our registered groups employed the latest version of VoroMQA[1] ("Voronoi diagram-based Model Quality Assessment"), our method for the estimation of protein structure quality that combines the idea of statistical potentials with the advanced use of the Voronoi tessellation of atomic balls.

Methods

Given a protein structure, it can be represented as a set of atomic balls, each ball having a van der Waals radius corresponding to the atom type. A ball can be assigned a region of space that contains all the points that are closer (or equally close) to that ball than to any other. Such a region is called a Voronoi cell and the partitioning of space into Voronoi cells is called Voronoi tessellation or Voronoi diagram. Two adjacent Voronoi cells share a set of points that form a surface called a Voronoi face. A Voronoi face can be viewed as a geometric representation of a contact between two atoms. The Voronoi cells of atomic balls may be constrained inside the boundaries defined by the solvent accessible surface (SAS) of the same balls. The procedure to construct the described surfaces is implemented as part of Voronota software[2]. It uses triangulated representations of Voronoi faces and spherical surfaces, contact areas are calculated as the areas of the corresponding triangulations.

In VoroMQA, inter-atomic and solvent contact areas are used to evaluate the quality of protein structural models by employing the idea of a knowledge-based statistical potential. The VoroMQA scoring function produces quality scores at different levels including atoms, residues and the full structure. The scoring function was not optimized or trained in any way to correspond to any reference based model quality-assessment scores: unsupervised learning was performed using experimentally determined structures of protein biological assemblies as input.

For CASP13 we attempted to improve VoroMQA by additionally considering hydrogen atoms (previously we used only heavy atoms). Reduce software tool[3] was employed for calculating coordinates of hydrogens. The resulting experimental VoroMQA version was run by two server groups: VoroMQA-A and VoroMQA-B. VoroMQA-A server preprocessed input models by rebuilding their side-chains using SCWRL4[4], VoroMQA-B did not alter input models before evaluating them.

The VoroMQA-select method used two versions of VoroMQA (with and without hydrogen atoms) and considered both unaltered and altered (processed with SCWRL4) variants of CASP13 server models. The calculated scores were used to construct a tournament-based ranking of models. VoroMQA was also used to determine if model structures contained unstructured N-terminal or C-terminal regions that needed to be removed prior to evaluation. This procedure was semi-automatic and required manual confirmation of trimming positions.

Availability

The VoroMQA web application is available at bioinformatics.ibt.lt/wtsam/voromqa. VoroMQA software for Linux is included in the Voronota package freely available from bitbucket.org/kliment/voronota/downloads.

1. Olechnovic, K., and Venclovas, C. (2017) VoroMQA: Assessment of protein structure quality using interatomic contact areas. *Proteins*, **85**(6), 1131-1145.
2. Olechnovic, K., and Venclovas, C. (2014) Voronota: a fast and reliable tool for computing the vertices of the Voronoi diagram of atomic balls. *J Comput Chem*, **35**(8), 672-681.
3. Word J.M., Lovell S.C., Richardson J.S., and Richardson D.C. (1999) Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation., **285**(4), 1735-47.
4. Krivov, G.G., Shapovalov, M.V., and Dunbrack, R.L. (2009) Improved prediction of protein side-chain conformations with SCWRL4. *Proteins*, **77**(4), 778-795.

Combining ProQ2 and Pcons to improve model quality assessment, selection and refinement

Björn Wallner

Linkӧping University, Dept. Physics, Chemistry, and Biology, Bioinformatics Division S-581 83 Linkӧping, Sweden

bjorn.wallner@liu.se

In the past, we have tried numerous ways to combine a single-model MQAP, ProQ2[1,2], with consensus based MQAP, Pcons[3], but in the end a simple linear combination is as good as any other more advanced combination. And it makes sense, the consensus methods usually fail to pick up the best models when there is no consensus and then score will be mostly based on the single-model MQAP, and when there is high consensus the single-model MQAP will select models among these. For our predictions in CASP13 we used Pcomb=0.2*ProQ2+0.8*Pcons. However, we did not limit ourselves to Pcomb but also manually inspected the top-ranked server models for ProQ2 and Pcons as well. Below, we describe our combined approach in the TS and TR categories.

Methods

The Wallner group participated in the TS, TR and QA categories in CASP13.

In the TS category we used ProQ2, Pcons and Pcomb to assess the local and global quality of all server models. The models, and in particular the per residue predicted distance deviation of for the top five models by ProQ2, Pcons and Pcomb were manually inspected. High quality regions, i.e regions with low predicted distance deviation to native were identified and restrained in a Rosetta relax simulation[4] using the predicted distance deviation. For each model,

For difficult targets with low consensus, defined by Pcons < 0.2, relaxed decoys were generated for 384h (16*24h) for all models. For easy targets, defined by Pcons >=0.2, relaxed decoys were generated for 384h (16*24h) for the first ranked and for 72h (3*24h) for rank 2-5. In total, 1,272,781 decoys were generated, ranging from 126 to 78,996 decoys per target depending in target size and target difficulty. A score that combined the Rosetta Energy and ProQ2 with equal weights were calculated (RosettaEnergy-ProQ2, minus since for ProQ2 higher is better). The five models with the best (lowest) combined score were selected and the ProQ2 local quality prediction was added to the B-factor column.

For the TR category, we used a similar approach as for the TS category, except for using the refinement starting structure instead of models top ranked by ProQ2, Pcons or Pcomb. We now manually inspected the predicted distance deviation by ProQ2, Pcons and Pcomb and chose a distance deviation cutoff that selected which part of the structure we should restraint. At most 16,000 decoys were generated for 384h (16*24h) using the coordinate constraint option in the relax protocol in Rosetta[4], resulting in 2,948 to 16,000 decoys per target depending on protein size. As above we calculated a combined Rosetta Energy and ProQ2 score and selected the five best models according to this score.

For the QA category the quality assessment is based on Pcomb only, i.e Pcomb=0.2*ProQ2+0.8Pcons. This is identical to what was used in CASP11 and CASP12

Availability

The current version of ProQ2 is currently available as a scoring function in the Rosetta modeling suite from http://www.rosettacommons.org. Additional scripts to generate all necessary input files can be found here: http://github.com/bjornwallner/ProQ_scripts/.

1. Ray, A., Lindahl, E. & Wallner, B. Improved model quality assessment using ProQ2. *BMC Bioinformatics* **13,** 224 (2012).
2 Uziela K, Wallner B. ProQ2: estimation of model accuracy implemented in Rosetta. *Bioinformatics* **32**(9):1411-3 (2016).
3. Wallner, B. & Elofsson, A. Pcons5: combining consensus, structural evaluation and fold recognition scores. *Bioinformatics* **21,** (2005).
4. P. Conway*, M. Tyka*, F. DiMaio*, D. Konerding and D. Baker (2013). Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein Science.*

CASP13 Tertiary Structure Prediction by wfAll-Cheng of the WeFold Collaborative

Jie Hou[1], Renzhi Cao[2], and Jianlin Cheng[1*]

1 - Department of Electrical Engineering & Computer Science, University of Missouri, Columbia, MO 65211, USA, 2 - Department of Computer Science, Pacific Lutheran University, WA 98447, USA,

*chengji@missouri.edu

WeFold[1] is a community-wide collaborative experiment for protein structure prediction. As part of the WeFold, we evaluated all the models provided by all WeFold branches and selected 5 top ranked models for CASP13 submission. The two main improvements benchmarked in the CASP13 are: (1) domain-based model evaluation applied to individual domains of multi-domain targets; and (2) large-scale model quality assessment empowered by deep learning and contact predictions.

Methods

Our wfAll-Cheng server firstly collected all the WeFold models and CASP13 server models for each target. A new de novo protein structure prediction pipeline in Wefold branches was used for generating protein decoys from protein sequence. During the prediction process, the step-wise conformational searching[2] was used instead of random sampling to generate thousands of decoys assisted with contact information generated by MetaPSICOV2[3]. We used Qprob[4] to rank all models and selected the best five models from all generated decoys as wfAll-Cheng-Cao models (for details, see our CASP13 abstract entitled "Collaborative de novo protein structure prediction using stepwise fragment sampling with help of contact prediction and model selection based on deep learning techniques"). Besides, the top 500 wfRosetta-Maghrabi models from around 10000 BAKER-ROSETTASERVER decoys were further reduced to 5 final models using Qprob[4] as wfAll-Cheng-Maghrabi models. The combined pool of all CASP13 server models, wfAll-Cheng-Cao and wfAll-Cheng-Maghrabi models for each target was evaluated by 13 complementary model quality metrics derived from contact predictions by DNCON2[5], single-model quality assessments (i.e. SBROD, OPUS_PSP[6], Model evaluator[7], RF_CB_SRS_OD[8], Rwplus[9], QMEAN[10] and Voronota[11]) and multi-model quality assessment (i.e. Pcons[12]). These quality scores were as input for our new deep learning method to generate a consensus ranking of models of each target (for details, see our CASP13 abstract entitled "Large-scale integration of protein model quality assessment using deep learning and contact prediction"). The top five models were selected as final prediction for wfAll-Cheng. If a protein was parsed into multiple domains, the models of individual domains were evaluated separately, and top 5 models of individual domains were combined into five final full-length models.

1 Keasar, C. *et al.* An analysis and evaluation of the WeFold collaborative for protein structure prediction and its pipelines in CASP11 and CASP12. *Scientific reports* **8**, 9939 (2018).

2 Bhattacharya, D., Cao, R. & Cheng, J. UniCon3D: de novo protein structure prediction using united-residue conformational search via stepwise, probabilistic sampling. *Bioinformatics* **32**, 2791-2799 (2016).

3 Buchan, D. W. & Jones, D. T. Improved protein contact predictions with the MetaPSICOV2 server in CASP12. *Proteins: Structure, Function, and Bioinformatics* **86**, 78-83 (2018).

4 Cao, R. & Cheng, J. Protein single-model quality assessment by feature-based probability density functions. *Scientific reports* **6**, 23990 (2016).

5 Adhikari, B., Hou, J. & Cheng, J. DNCON2: Improved protein contact prediction using two-level deep convolutional neural networks. *Bioinformatics* (2017).

6 Lu, M., Dousis, A. D. & Ma, J. OPUS-PSP: an orientation-dependent statistical all-atom potential derived from side-chain packing. *Journal of molecular biology* **376**, 288-301 (2008).

7 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** (2009).

8 Rykunov, D. & Fiser, A. Effects of amino acid composition, finite size of proteins, and sparse statistics on distance-dependent statistical pair potentials. *Proteins: Structure, Function, and Bioinformatics* **67**, 559-568 (2007).

9 Zhang, J. & Zhang, Y. A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. *PloS one* **5**, e15386 (2010).

10 Benkert, P., Tosatto, S. C. & Schomburg, D. QMEAN: A comprehensive scoring function for model quality assessment. *Proteins: Structure, Function, and Bioinformatics* **71**, 261-277 (2008).

11 Olechnovič, K. & Venclovas, Č. Voronota: A fast and reliable tool for computing the vertices of the Voronoi diagram of atomic balls. *Journal of computational chemistry* **35**, 672-681 (2014).

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

Prediction of protein structure with the improve UNRES force field aided by contact prediction derived from evolutionarily related proteins

E.A. Lubecka[1,2*], A.G. Lipska[2], A.K. Sieradzan[2], A. Liwo[2], I. Anishchanka[3], D. Kim[3] and S.N. Crivelli[4]

1 - Institute of Informatics, Faculty of Mathematics, Physic and Informatics, University of Gdańsk, Wita Stwosza 57, 80-308 Gdańsk, Poland, 2 - Faculty of Chemistry, University of Gdańsk, Wita Stwosza 63, 80-308 Gdańsk, Poland, 3 - Department of Biochemistry, and Institute for Protein Design, University of Washington, Seattle, WA 98195, 4 - Department of Computer Science, UC Davis, One Shields Ave., Davis, CA 95616

emilia.lubecka@ug.edu.pl

Empirical force fields, both all-atom and coarse-grained, are still not capable of producing good models of proteins without external information. One source of this information are residue-residue contacts predicted based on proteins with known structures, which are evolutionarily related to the target protein. The purpose of this study was to assess how the use of predicted contacts can improve protein models obtained with the coarse-grained physics-based united-residue (UNRES)[1] force field, which was also used by the UNRES group in ab initio prediction mode.

Methods

The UNRES model[1] is designed to treat proteins. In this model, a polypeptide chain is represented as a sequence of α-carbon (Cα) atoms connected by virtual bonds, with united side chains (SC) attached to them and peptide groups (p) positioned in the middle between the two consecutive Cαs. The SCs and ps are the only interacting sites, while the Cαs only assist in geometry definition. The solvent is implicit in the force field. The same improved version of the force field as that used by the UNRES group was implemented[2], which has enhanced capacity of modeling beta-sheet structures. We have used Multiplexed Replica Exchange Molecular Dynamics (MREMD)[1] to run production simulations. The contact information was included in the potential energy as a sum of flat-bottom bounded penalty functions derived from the Lorentzian function, with barrier height proportional to contact probability[3]. The results were analyzed by Weighted-Histogram Analysis Method (WHAM) to calculate relative free energy of each structure of last slice of the MREMD simulation[1] and next a cluster analysis was employed to cluster the structures from a MREMD simulation. 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 by using the PULCHRA[4] and SCWRL[5] protocols and then refined by performing short restrained MD runs (restraints coming from the parent UNRES structures) by using the AMBER14 [6] with the ff14SB force field and implicit GBSA solvation model to give the models which were subsequently submitted.

Contact prediction, from which the distance restraints were derived, were carried out with the GREMLIN[7] method. GREMLIN works by constructing a global statistical model that simultaneously captures the conservation and co-evolution patterns in the input multiple sequence alignment. The alignments were generated using HHblits[8] and Jackhammer[9] with varying e-value and number of iterations. Strongly co-evolving residue pairs as identified by this approach, were used as restraint in modeling. For simple proteins, we have generated contacts directly from servers models. Additionally, for larger proteins, restraints from secondary structure prediction by PSIPRED[10] were imposed on the virtual-bond geometry.

All types of 3D-structure predictions were run (regular, data-assisted, and refinement). In data-assisted predictions (using SAXS/SANS, cross-linking, and simulated and real NMR data), appropriate penalty terms were added to the target functions. A special penalty function was engineered to handle ambiguous NMR restraints.

Results

We postpone the assessment of the approach until the official release of CASP13 results.

Availability

The UNRES package is available at www.unres.pl; GREMLIN at http://gremlin.bakerlab.org.

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, Chapter 8, pp. 107-122.
2. Sieradzan,A.K., Makowski,M., Augustynowicz,A. & Liwo,A. (2017) A general method for the derivation of the functional forms of the effective energy terms in coarse-grained energy functions of polymers. I. Backbone potentials of coarse-grained polypeptide chains. *J. Chem. Phys.* **146**, 124106.
3. Sieradzan,A.K. & Jakubowski,R. (2017) Introduction of steered molecular dynamics into UNRES coarse-grained simulations package. *J. Comput. Chem.* **38**, 553-562.
4. Rotkiewicz,P. & Skolnick,J. (2008) Fast procedure for reconstruction of full-atom protein models from reduced representations. *J. Comput. Chem.*, **29**, 1460-1465.
5. Wang,Q., Canutescu,A.A. & Dunbrack,R.L. (2008) SCWRL and MolIDE: Computer programs for side-chain conformation prediction and homology modeling. *Nat. Protoc.* ***3***,1832-1847.
6. Case,D.A., Babin,V., Berryman,J.Y., Betz,R.M., Cai,Q., Cerutti,D.S. et al. (2014) *AMBER 14*, University of California, San Francisco.
7. Kamisetty,H., Ovchinnikov,S. & Baker,D. (2013) Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 16674-16679.
8. 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.
9. Eddy,S.R. (2009) A new generation of homology search tools based on probabilistic inference. *Genome Inform.*, **23**, 205-211.
10. McGuffin,L.J., Bryson,K. & Jones,D.T. (2000) The PSIPRED protein structure prediction server. *Bioinformatics*, **16**, 404-405.

Tertiary Structure Prediction by the wfRosetta-ModF7 group

L.J. McGuffin[1], S.N. Crivelli[2] and A.H.A. Maghrabi[1]

1 - School of Biological Sciences, University of Reading, Reading, UK, 2 - Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, USA

l.j.mcguffin@reading.ac.uk

WeFold[1] is an open collaboration initiative for protein structure prediction within CASP. It brings together labs and individuals through the science gateway (http://wefold.nersc.gov/) and provides computing and storage resources through the National Energy Research Scientific Computing (NERSC) center. WeFold enables interaction among groups that work on different components of the protein structure prediction pipeline, thus making it possible to leverage expertise across prediction categories. The combination of these components creates hybrid protein structure prediction pipelines, each submitting its own models. This collaboration aims to promote a synergistic effect among the participants and ultimately produce better results than those achieved by the individual methods.

As part of the WeFold collaborative initiative for CASP13, we have tested our new pipeline, wfRosetta-ModF7, which combines the output from Rosetta[2], VoroMQA[3] and ModFOLD7.

Methods

Our group made use of the data provided by WeFOLD members, via the web-based forum, with contributors from different institutions. A huge number of protein domain decoys were generated and then uploaded to the WeFOLD forum for each CASP target. In our group, we focused on the analysis of the Rosetta server decoys only. These domain decoys were filtered, quality assessed and then recombined to full length models using the steps outlined below.

Filtering decoys:

In the first step, a large number of protein decoys were generated and provided to the WeFOLD forum by the Rosetta server[2] (see other abstract/s for details). Each protein target domain had up to 10000 generated Rosetta decoys. In the filtering stage, the number of Rosetta decoys were reduced from 10000 to 500 using the VoroMQA[3] global score rankings.

Model quality assessment:

After filtering, the remaining top 500 decoys were evaluated using the new version of our model quality assessment program - ModFOLD7_rank (see our ModFOLD7 abstract for details). The ModFOLD7 server has been updated from ModFOLD6 [4] which participated with CASP12 and performed well in the Estimates of Model Accuracy category[5].

Domain recombination and final model selection:

For the targets with more than one predicted domain, the top 5 ranked models for each domain were recombined using the domain_assembly script obtained from the Cheng group (Jie Hou, personal communication), which implements MODELLER[6]. The final set of full chain models were then scored and re-ranked again using ModFOLD7_rank. The per-residue error estimates were added to the B-factor column in each model file and the top 5 ranked models were submitted.

1. Keasar,C., McGuffin,L. J., Wallner,B., Chopra,G., Adhikari,B., Bhattacharya,D., Blake,L., Bortot,L.O., Cao,R., Dhanasekaran,B.K., Dimas,I., Faccioli,R.A., Faraggi,E., Ganzynkowicz,R., Ghosh,S., Ghosh,S., Giełdoń,A., Golon,L., He,Y., Heo,L., Hou,J., Khan,M., Khatib,F., Khoury,G.A., Kieslich,C., Kim,D.E., Krupa,P., Lee,G.R.,

Li,H., Li,J., Lipska,A., Liwo,A., Maghrabi,A.H.A., Mirdita,M., Mirzaei,S., Mozolewska,M.A., Onel,M., Ovchinnikov,S., Shah,A., Shah,U., Sidi,T., Sieradzan,A.K., Ślusarz,M., Ślusarz,R., Smadbeck,J., Tamamis,P., Trieber,N., Wirecki,T., Yin,Y., Zhang,Y., Bacardit,J., Baranowski,M., Chapman,N., Cooper,S., Defelicibus,A., Flatten,J., Koepnick,B., Popović,Z., Zaborowski,B., Baker,D., Cheng,J., Czaplewski,C., Delbem,A.C.B., Floudas,C., Kloczkowski,A., Ołdziej,S., Levitt,M., Scheraga,H., Seok,C., Söding,J., Vishveshwara,S., Xu,D. and Crivelli,S.N. (2018) An analysis and evaluation of the WeFold collaborative for protein structure prediction and its pipelines in CASP11 and CASP12. *Scientific Reports*, **8**. 9939.

2. Kim,D.E., Chivian,D., & Baker,D. (2004). Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res.* **32**, W526-W531.
3. Olechnovič,K. & Venclovas,Č. (2017) VoroMQA: Assessment of protein structure quality using interatomic contact areas. *Proteins*. **85**, 1131-1145.
4. Maghrabi,A.H.A. & McGuffin,L.J. (2017) ModFOLD6: an accurate web server for the global and local quality estimation of 3D models of proteins. *Nucleic Acids Res.* **45**, W416-W421.
5. Kryshtafovych,A., Monastyrskyy,B., Fidelis,K., Schwede,T. & Tramontano,A. (2018) Assessment of model accuracy estimations in CASP12. *Proteins*. **86 S1**, 345-360.
6. Webb,B. & Sali,A. (2016) Comparative Protein Structure Modeling Using Modeller. *Current Protocols in Bioinformatics.* **54**, John Wiley & Sons, Inc., 5.6.1-5.6.37.

CASP13 Tertiary Structure Prediction by the wfRosetta-PQ2-AngQA group

Renzhi Cao[1*], David E Kim[2], Silvia Crivelli[3]

1 - Department of Computer Science, Pacific Lutheran University, Tacoma, WA 98447; 2 - University of Washington, Seattle, WA; 3 - Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA;

*caora@plu.edu

In CASP 13, we have tested our new pipeline as part of WeFold[1], which is a web-based effort to foster collaboration environment within CASP community. The protein decoys are generated and uploaded to WeFold web server to share between different institutions, and the top 5 protein decoys are selected from all decoys as the final submission. We have tested two different novel protein model Quality Assessment (QA) methods - ProQ[2] and AngularQA(details of the AngularQA method is described in CASP13 abstract entitled "AngularQA: Protein Model Quality Assessment with LSTM Networks") in this pipeline.

Methods

First of all, the protein decoys are generated by Rosetta group[3] for each protein sequence. This step may take several days. The decoys will be uploaded to WeFold server after they are generated. We are going to download all decoys from the WeFold server. There could be more than 10,000 decoys for each protein target, so model selection is very important step. There are two novel QA methods used in this pipeline. The first method is ProQ2, which is state-of-the-art QA methods used the machine learning method - Support Vector Machine (SVM) for model quality assessment. It would finally select a small portion of decoys out of all generated decoys. After the reduced decoys are uploaded to WeFold server, the second method AngularQA is going to be applied to those protein decoys. It is a new QA method that analyze protein decoys with a novel representation and first time applied LSTM network for model ranking. The AngularQA method will be applied to decoys reduced by ProQ2, and finally selected five decoys as the final prediction.

1. Keasar, C., Foldit Players consortium, McGuffin, L.J., Wallner, B., Chopra, G., Adhikari, B., Bhattacharya, D., Blake, L., Bortot, L.O., Cao, R., Dhanasekaran, B.K., Dimas, I., Faccioli, R.A., Faraggi, E., Ganzynkowicz, R., Ghosh, S., Ghosh, S., Giełdoń, A., Golon, L., He, Y., Heo, L., Hou, J., Khan, M., Khatib, F., Khoury, G.A., Kieslich, C., Kim, D.E., Krupa, P., Lee, G.R., Li, H., Li, J., Lipska, A., Liwo, A., Maghrabi, A.H.A., Mirdita, M., Mirzaei, S., Mozolewska, M.A., Onel, M., Ovchinnikov, S., Shah, A., Shah, U., Sidi, T., Sieradzan, A.K., Ślusarz, M., Ślusarz, R., Smadbeck, J., Tamamis, P., Trieber, N., Wirecki, T., Yin, Y., Zhang, Y., Bacardit, J., Baranowski, M., Chapman, N., Cooper, S., Defelicibus, A., Flatten, J., Koepnick, B., Popović, Z., Zaborowski, B., Baker, D., Cheng, J., Czaplewski, C., Delbem, A.C.B., Floudas, C., Kloczkowski, A., Ołdziej, S., Levitt, M., Scheraga, H., Seok, C., Söding, J., Vishveshwara, S., Xu, D. & Crivelli, S.N. (2018). An analysis and evaluation of the WeFold collaborative for protein structure prediction and its pipelines in CASP11 and CASP12. *Sci. Rep.* **8**,
2. Ray, A., Lindahl, E. & Wallner, B. (2012). Improved model quality assessment using ProQ2. *BMC Bioinformatics* **13**, 224
3. Ovchinnikov, S., Park, H., Kim, D.E., DiMaio, F. & Baker, D. (2018). Protein structure prediction using Rosetta in CASP12. *Proteins* **86 Suppl 1**, 113–121

CASP13 Tertiary Structure Prediction by the wfRstta-Maghrabi-TQA group

Renzhi Cao[1*], Ali Hassan Ahmed Maghrabi[2], Silvia Crivelli[3]

1 - Department of Computer Science, Pacific Lutheran University, Tacoma, WA 98447; 2 - College of Applied Sciences, Umm al-Qura University, Mecca, Saudi Arabia; 3 - Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA;

*caora@plu.edu

In CASP 13, we have tested our new pipeline as part of WeFold[1], which is a web-based effort to foster collaboration environment within CASP community. The protein decoys are generated and uploaded to WeFold web server to share between different institutions, and the top 5 protein decoys are selected from all decoys as the final submission. We have tested two different novel protein model Quality Assessment (QA) methods - VoroMQA[2] and TopQA(details of this method is described in CASP13 abstract entitled "TopQA: A Topological Representation for Single-Model Protein Quality Assessment with Machine Learning Technique") in this pipeline.

Methods

First of all, Rosetta group provided the protein decoys generated by Rosetta[3] for each protein sequence. The decoys were uploaded to WeFold server after generated. There would be 10,000 decoys for each protein target, so model selection is very important step. For the model selection part, there are two novel QA methods used in this pipeline. The first method is VoroMQA, which did assessment of protein structure quality using interatomic contact area. It would finally select 500 decoys out of 10,000 decoys. The second method is TopQA. It is a new method that analyze protein decoys from the topological representation instead of features that traditional QA methods usually used. The TopQA method will be applied to 500 decoys reduced by VoroMQA, and finally selected five decoys as the final prediction.

1. Keasar, C., Foldit Players consortium, McGuffin, L.J., Wallner, B., Chopra, G., Adhikari, B., Bhattacharya, D., Blake, L., Bortot, L.O., Cao, R., Dhanasekaran, B.K., Dimas, I., Faccioli, R.A., Faraggi, E., Ganzynkowicz, R., Ghosh, S., Ghosh, S., Giełdoń, A., Golon, L., He, Y., Heo, L., Hou, J., Khan, M., Khatib, F., Khoury, G.A., Kieslich, C., Kim, D.E., Krupa, P., Lee, G.R., Li, H., Li, J., Lipska, A., Liwo, A., Maghrabi, A.H.A., Mirdita, M., Mirzaei, S., Mozolewska, M.A., Onel, M., Ovchinnikov, S., Shah, A., Shah, U., Sidi, T., Sieradzan, A.K., Ślusarz, M., Ślusarz, R., Smadbeck, J., Tamamis, P., Trieber, N., Wirecki, T., Yin, Y., Zhang, Y., Bacardit, J., Baranowski, M., Chapman, N., Cooper, S., Defelicibus, A., Flatten, J., Koepnick, B., Popović, Z., Zaborowski, B., Baker, D., Cheng, J., Czaplewski, C., Delbem, A.C.B., Floudas, C., Kloczkowski, A., Ołdziej, S., Levitt, M., Scheraga, H., Seok, C., Söding, J., Vishveshwara, S., Xu, D. & Crivelli, S.N. (2018). An analysis and evaluation of the WeFold collaborative for protein structure prediction and its pipelines in CASP11 and CASP12. *Sci. Rep.* **8**,
2. Olechnovič, K. & Venclovas, Č. (2017). VoroMQA: Assessment of protein structure quality using interatomic contact areas. *Proteins* **85**, 1131–1145
3. Ovchinnikov, S., Park, H., Kim, D.E., DiMaio, F. & Baker, D. (2018). Protein structure prediction using Rosetta in CASP12. *Proteins* **86 Suppl 1**, 113–121

Protein contact prediction and contact-assisted threading

Shuo Pan#, Qi Wu#, Jianyi Yang

School of Mathematical Sciences, Nankai University, Tianjin, 300071, China (# co-first authors).

yangjy@nankai.edu.cn

In CASP13, the protein contacts were predicted using multiple sequence alignment (MSA) from multiple sequence databases and two-stage deep residual neural networks (Yang-Server and RRMD). The predicted contacts are then used to develop a new contact-assisted threading algorithm, which was incorporated into the I-TASSER Suite [1] to improve the step of template identification (Yang-Server and CMA-align).

Methods

It turns out that the number of homologous sequences is one of the most key factors affecting the performance of contact prediction methods. The recent work from the Baker group suggests that the metagenome sequences can be used to build deeper MSA, making it possible to generate contact maps and structure models for 614 protein families that have no known structures [2]. We aim to explore more about the usage of metagenome sequences in the predition of protein contact and structure.

For contact prediction of each target, an MSA was generated using multiple sequence databases, including Uniclust30 [3], UNIREF100 and metagenome50 [4]. Two-stage deep residual neural networks were used to predict the contacts. For each residue pair, the inputs of the first stage are 231 covariance features, a reduced set from the work of DeepCov [5]. The inputs of the second stage include predicted secondary structure and relative solvent accessibility, and two predicted contact maps from the first stage and the direct-coupling method CCMpred [6], respectively.

For structure prediction, the predicted contact map from the second stage was used for template detection by a two-step dynamic programming. In the first step, an initial alignment between the query and each template was generated with profile-profile alignment, in which the profile includes the position-specific scoring matrix and the predicted secondary structure and relative solvent accessibility. The initial alignment was refined iteratively in the second step by matching the predicted contact map of the query and the native contact map of each template. The top templates detected from the second step were added into I-TASSER Suite's template pool for building full-length atomic structure models. The contact map was also used to re-rank the predicted structure models, when the map was predicted with high confidence as judged by the number of homologous sequences and the probability scores of top contacts.

Results

In our preliminary benchmark tests with 38 hard targets from CASP12, the top L/5 precision of the predicted contact maps by our method is about 10% higher than DeepCov and comparable to or higher than SPOT-Contact [7] and RaptorX-Contact [8]. The TM-scores of the predicted structures for 83 CASP12 targets by our method are consistently higher than LOMETS [9], HHsearch [10] and EigenTHREADER [11].

1. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: protein structure and function prediction. *Nature methods* 2015;12(1):7-8.
2. Ovchinnikov S, Park H, Varghese N, Huang PS, Pavlopoulos GA, Kim DE, Kamisetty H, Kyrpides NC, Baker D. Protein structure determination using metagenome sequence data. Science 2017;355(6322):294-298.
3. Mirdita M, von den Driesch L, Galiez C, Martin MJ, Soding J, Steinegger M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic acids research 2017;45(D1):D170-D176.
4. Steinegger M, Soding J. Clustering huge protein sequence sets in linear time. Nature communications

2018;9(1):2542.

5. Jones DT, Kandathil SM. High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. Bioinformatics 2018.
6. Seemayer S, Gruber M, Söding J. CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. Bioinformatics 2014;30(21):3128-3130.
7. Wang S, Sun S, Li Z, Zhang R, Xu J. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. PLOS Computational Biology 2017;13(1):e1005324.
8. Hanson J, Paliwal K, Litfin T, Yang Y, Zhou Y. Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks. Bioinformatics 2018:bty481-bty481.
9. Wu S, Zhang Y. LOMETS: a local meta-threading-server for protein structure prediction. Nucleic acids research 2007;35(10):3375-3382.
10. Soding J. Protein homology detection by HMM-HMM comparison. Bioinformatics 2005;21(7):951-960.
11. Buchan DWA, Jones DT. EigenTHREADER: analogous protein fold recognition by efficient contact map threading. Bioinformatics 2017;33(17):2684-2690.

The YASARA homology modeling module V4.2 with new profile and threading methods for remote template recognition

E. Krieger

YASARA Biosciences GmbH, Wagramer Strasse 25/3/45, 1220 Vienna, Austria

elmar@yasara.org

Up to CASP 11, the **YASARA Structure** server (www.yasara.org/homologymodeling) submitted predictions only for classic homology modeling targets, where template identification was easy, since high-resolution homology modeling including ligands is one of the main applications of YASARA. For CASP12, remote fold recognition methods were developed, which often helped to identify useful templates, but equally often failed. Since CASP ranking is usually based on summed up GDT_TS scores, also the failures were submitted, which would however not be used in practice, since they are automatically classified as trash in YASARA's homology modeling report.

Methods

As in previous CASPs[1], our method targets homology modeling with a focus on high-resolution refinement, new folds can hardly be predicted. First PsiBLAST is run with Uniref90 profiles to identify potential templates. If no reliable hit is found, the remote fold recognition procedure is started, which scans a library of ~60000 representative PDB structures using a sensitive profile-profile alignment with the Smith&Waterman algorithm. The match between PSI-Predicted[2] and actual secondary structure is included in the score, the top hits are validated by building fast approximate all-atom models, to make sure that the aligned fragments pack together well.

High-resolution models are built for the top 10 templates, using stochastic[3] profile-profile alignments including SSALN features[4] to arrive at up to five alternative high-scoring target-template alignments, building models for all of them (using SCWRL[5] rotamer libraries, but additional energy terms), and scoring them. The best parts of the up to 50 models are fused to form a hybrid model. 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.

The best-scoring model was subjected to high-resolution refinement, running 100 short MD simulations in parallel with the partly knowledge-based YASARA force field[1] and newly developed high-performance simulation algorithms[6]. The best refined model was submitted as model 1, the unrefined model as model 2, and models 3 to 5 were based on alternative alignments and/or alternative templates.

Results

The recipe above yielded homology models for essentially all CASP13 targets, many of which where classified as trash, but submitted anyway for reasons explained above. The whole procedure was implemented as a fully automatic server, requiring human intervention only for occasional bug fixes.

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. Jones, D.T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. J.Mol.Biol. 292, 195-202
3. Mueckstein, U., Hofacker, I.L. and Stadler, P.F. (2002). Stochastic pairwise alignments. Bioinformatics 18 Sup2, 153-160
4. 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
5. Canutescu, A.A., Shelenkov, A.A. and Dunbrack, R.L. Jr. (2003). A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci. 12, 2001-2014.
6. Krieger E, Vriend G (2015). New ways to boost molecular dynamics simulations. J.Comput.Chem. 36, 996-1007

Protein tertiary structure predictions by Zhang human group in CASP13

Wei Zheng[1], Chengxin Zhang[1], Yang Li[1,2], S M Golam Mortuza[1], Yang Zhang[1]

1- Department of Computational Medicine and Bioinformatics, Department of Biological Chemistry, University of Michigan, 100 Washtenaw Ave, Ann Arbor, MI 48109; 2- School of Computer Science and Engineering, Nanjing University of Science and Technology, Xiaolingwei 200, Nanjing, China, 210094

yangzhanglab@umich.edu

The tertiary structure prediction of the Zhang human group in CASP13 is based on the C-I-TASSER pipeline that is identical to that used in the Zhang-Server group (see Zhang-Server Abstract), except that the whole set of structure models generated by the CASP servers, instead of the in-house LOMETS templates,[1] were used as the starting models of the C-I-TASSER pipeline. The sequence-based contact restraints are generated by two in-house contact-map predictors (NeBcon[2] and ResPRE), which are used for guiding the structural assembly simulations. One purpose of our participating in the human section is to examine the impact of different initial threading templates on the final models of the C-I-TASSER pipeline.

Similar to the Zhang-Server pipeline, for the query proteins that were deemed by LOMETS[1] as Hard and Very-Hard targets, the models generated by C-QUARK (Mortuza et al, in preparation) are used to sort and re-rank the CASP server models based on their TM-score to the *ab initio* folding models, under the hypothesis that a close match of the models from *ab initio* folding and template-based modeling is significant and often indicates the correctness of the folds. The consensus restraints are then extracted from both *ab initio* structural models and the CASP server models, which are used to guide the C-I-TASSER structure assembly simulations. In addition, the sequence-based contact maps, generated NeBcon[2] and a recently developed deep-learning based contact predictor, ResPRE (Li et al, in preparation), are incorporated into the C-I-TASSER force field. The final models were selected by SPICKER[3] from the simulation trajectories, which are further refined at atomic level by the fragment-guided molecular dynamic (FG-MD) simulations.[4]

For multiple-domain proteins, ThreaDom[5] was used to predict the domain boundary and linker regions from the LOMETS threading alignments.[1] Full-length models are assembled from the individual domain structures using a rigid-body domain docking and assembly algorithm, DEMO (Zhou et al, in preparation). The procedure is fully automated.

1. Wu, S.; Zhang, Y., LOMETS: A local meta-threading-server for protein structure prediction. *Nucl. Acids. Res.* 2007, **35**, 3375-3382.
2. He, B.; Mortuza, S. M.; Wang, Y.; Shen, H. B.; Zhang, Y., NeBcon: protein contact map prediction using neural network training coupled with naive Bayes classifiers. *Bioinformatics* 2017, **33**, 2296-2306.
3. Zhang, Y.; Skolnick, J., SPICKER: A clustering approach to identify near-native protein folds. *J Comput Chem* 2004, **25**, 865-71.
4. Zhang, J.; Liang, Y.; Zhang, Y., Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling. *Structure* 2011, **19**, 1784-95.
5. Xue, Z.; Xu, D.; Wang, Y.; Zhang, Y., ThreaDom: extracting protein domain boundary information from multiple threading alignments. *Bioinformatics* 2013, **29**, i247-i256.

CEthreader: Detecting distant-homology proteins using contact map guided threading

Wei Zheng[1], Qiqige Wuyun[2] and Yang Zhang[1]

1 -Department of Computational Medicine and Bioinformatics, Department of Biological Chemistry, University of Michigan, 100 Washtenaw Ave, Ann Arbor, MI 48109, 2 - Computer Science and Engineering Department, Michigan State University, East Lansing, MI 48823 USA

yangzhanglab@umich.edu

We developed a threading method, called CEthreader (Contact Eigenvector-based threader), which converts contact map into a set of single-body Eigenvectors through Eigendecomposition. CEthreader subsequently performs dynamic programming based on the contact Eigenvectors, secondary structure and sequence profile to identify templates.

Methods

Step1 features preparation. In this step, an in-house MSA (multiple sequence alignment) construction software collects homology sequences from UNICLUST[1], UNIREF90[2] and METACLUST[3] databases. The MSA is used in building Henikoff[4] profile and in secondary structure prediction by PSSpred. The residue-residue contacts with sequence separation of ≥ 5 are predicted by NeBcon[5]. Besides, the residue pairs located at the same helix with a sequence separation of 4 are also taken into account. The residue pairs are ranked in descending order of the confidence score of the predicted contacts by NeBcon. The top 2.5L residue pairs that are in contacts are selected to form the final contact map of the query.

Step 2 templates detection. For a given protein with the length L, its contact map is an L×L square, binary and symmetric matrix, where residues that are in contacts (Cβ distance < 8 Å) are designated as 1 and non-contacting residues are set as 0. Eigendecomposition of the contact map is performed, followed by selection of the largest K Eigenvalues and corresponding Eigenvectors to calculate the K-dimensional contact Eigenvector sequences. A scoring function combining contact Eigenvectors, secondary structure and profile terms is utilized by semi-global dynamic programming to compare the query sequence with each template in our database. Then different templates are ranked by contact map overlap between query and template. Finally, top 5 templates are selected for building full length models.

Step 3 models generation. The final models are built by MODELLER[6] with restraints generated from templates and predicted secondary structure. The pairwise TM-score[7] of 5 templates will be calculated. If the average TM-score is less than 0.85, 5 models are built from 5 templates individually. Otherwise, we build model 1 using restraints from all 5 templates, model 2 using restraints from top 4 templates, and so on. Finally, 5 models are submitted.

1. Mirdita, M., von den Driesch, L., Galiez, C., Martin, M. J., Söding, J. & Steinegger, M. (2017). Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Research* **45**, D170-D176.
2. Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H. & the UniProt, C. (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926-932.
3. Steinegger, M. & Söding, J. (2018). Clustering huge protein sequence sets in linear time. *Nature Communications* **9**, 2542.

4. Henikoff, S. & Henikoff, J. G. (1994). Position-based sequence weights. *Journal of Molecular Biology* **243**, 574-578.
5. He, B., Mortuza, S. M., Wang, Y., Shen, H.-B. & Zhang, Y. (2017). NeBcon: protein contact map prediction using neural network training coupled with naïve Bayes classifiers. *Bioinformatics* **33**, 2296-2306.
6. Šali, A. & Blundell, T. L. (1993). Comparative Protein Modelling by Satisfaction of Spatial Restraints. *Journal of Molecular Biology* **234**, 779-815.
7. Xu, J. & Zhang, Y. (2010). How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **26**, 889-95.

NeBcon: Contact map prediction using neural network training coupled with naïve Bayes classifiers

S.M. Mortuza[1], Yang Zhang[1]

1 Department of Computational Medicine and Bioinformatics, Department of Biological Chemistry, University of Michigan, 100 Washtenaw Ave, Ann Arbor, MI 48109

yangzhanglab@umich.edu

We extended our previously developed pipeline, NeBcon,[1] where naïve Bayes classifier (NBC) was used to combine the contact map predictions from six predictors. The posterior probabilities from the NBC model were further trained along with structural and coevolution features through neural network to obtain final contact map models.

Methods

In order to predict contact map for each query sequence, the NeBcon pipeline performed three steps: i) multiple sequence alignment (MSA) generation, ii) contact map predictions from six predictors using the MSA, and calculation of posterior probability from the predictions using NBC, and iii) final contact map prediction from neural network training of the NBC probabilities and structural and coevolution features.

The MSA for each query sequence was derived by HHblits[2] search against Uniclust30[3] database, and jackhammer[4] search against UniRef90[5]. The MetaClust metagenome sequence database[6] search was performed using hmmsearch[7] in order to enrich the MSA when the number of effective sequence (*Neff)* is lower than 128.

The generated MSA was used to predict contacts from ResPRE, DeepPLM, DeepCov[8], DNCON2[9], MetaPSICOV2[10] and GREMLIN[11], where ResPRE and DeepPLM are two in-house contact predictors based on deep convolutional neural network (CNN) models[12]. The predictions from the six predictors were combined using NBC, which are built on the posterior probabilities of residues to be in contact given the confidence scores of the predictions from the individual predictors.

The NBC model was further coupled with neural network models, which were trained on probabilities of the Bayes combination, and structural and coevolution features, including secondary structure, solvent accessibility, Shannon entropy, *Neff*, residue composition, contact potential, and mutual information. The coevolution features were extracted from the generated MSA. The classification of contacting and non-contacting residues in the neural network training was based on three thresholds of $C\beta$- $C\beta$ distance (7Å, 8Å and 9 Å) of residues in the dataset comprising 1,066 non-redundant proteins. In order to utilize a bulk portion of the dataset for each distance threshold that has significant imbalance between the number of contacting (positive sample) and non-contacting (negative sample) long-range residues pairs, we selected five sets of samples; each comprising approximately 2:23 ratio of contacting and non-contacting long-range pairs. While same positive samples were taken, the negative samples were non-redundant in each set. Outputs from five neural network models trained on the five sets of samples were averaged for each of the distance threshold. The average results for the three distance thresholds were further averaged to obtain the final contact map of the query sequence.

Availability

The web server of NeBcon is available at https://zhanglab.ccmb.med.umich.edu/NeBcon.

1. He,B. J.; Mortuza,S. M.; Wang,Y. T.; Shen,H. B.; Zhang,Y., NeBcon: protein contact map prediction using neural network training coupled with naiive Bayes classifiers. *Bioinformatics* **2017,** *33* (15), 2296-2306.
2. Remmert,M.; Biegert,A.; Hauser,A.; Soding,J., HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods* **2012,** *9* (2), 173-175.
3. Mirdita,M.; von den Driesch,L.; Galiez,C.; Martin,M. J.; Soding,J.; Steinegger,M., Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Research* **2017,** *45* (D1), D170-D176.
4. Johnson,L. S.; Eddy,S. R.; Portugaly, E., Hidden Markov model speed heuristic and iterative HMM search procedure. *Bmc Bioinformatics* **2010,** *11*.
5. Suzek,B. E.; Wang,Y. Q.; Huang,H. Z.; McGarvey,P. B.; Wu,C. H.; UniProt, C., UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **2015,** *31* (6), 926-932.
6. Steinegger,M.; Söding,J., Clustering huge protein sequence sets in linear time. *Nature Communications* **2018,** *9* (1), 2542.
7. Eddy,S. R., Accelerated Profile HMM Searches. *Plos Computational Biology* **2011,** *7* (10).
8. Jones,D. T.; Kandathil,S. M., High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. *Bioinformatics* **2018**, bty341-bty341.
9. Adhikari,B.; Hou,J.; Cheng,J., DNCON2: improved protein contact prediction using two-level deep convolutional neural networks. *Bioinformatics* **2017,** (btx781).
10. Buchan,D. W. A.; Jones,D. T., Improved protein contact predictions with the MetaPSICOV2 server in CASP12. *Proteins-Structure Function and Bioinformatics* **2018,** *86*, 78-83.
11. Kamisetty,H.; Ovchinnikov,S.; Baker, D., Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proceedings of the National Academy of Sciences of the United States of America* **2013,** *110* (39), 15674-15679.
12. Li,Y.; Yu,D.; Zhang,Y., ResPRE: High-accuracy protein contact map prediction by integrating precision matrix with deep residual neural networks. **2018**.

A meta-approach to atomic-level protein structure refinement

Wei Zheng[1] and Yang Zhang[1]

1 -Department of Computational Medicine and Bioinformatics, Department of Biological Chemistry, University of Michigan, 100 Washtenaw Ave, Ann Arbor, MI 48109

yangzhanglab@umich.edu

We participate in CASP13 structure refinement experiment as a human group "Zhang-Refinement", with a new refinement approach, which utilizes seven individual protein structure refinement programs to generate decoys, and then pick up models by SPICKER[1] and MolProbity score[2].

Methods

Templates detection. From the initial model released by CASP, a collection of homology templates with a TM-score[3]>0.5 is identified by TM-align[4] from a non-redundant PDB database. The top 10 templates ranked by TM-score will be selected. For each selected template, a Cα distance map is collected from the TM-align aligned regions. Compared to the distance map of the initial model, we remove all restraints that satisfy $\left|d_{i,j}^{template} - d_{i,j}^{initial}\right| > 8\ \text{Å}$ from the template distance maps. Then a multiple Gaussian distribution style (EQ1) energy potential is used as FG-MD[5] and ModRefiner-TBM distance restraints.

$$P_{distance-restraint} = -\sum_{i,j}^{L} \ln \frac{\sum_{template\_k=1}^{10} \frac{w_{template\_k}}{\sqrt{2\pi}\sigma} e^{\frac{-(d-d_{i,j}^{template\_k})}{2\sigma^2}}}{P_{background}} \qquad EQ1$$

Decoys generation. In this step, we utilize 7 individual refinement approaches to generate decoys, including FG-MD, ModRefiner[6], ModRefiner-TBM, GalaxyRefine[7], i3DRefine[8], MESHI[9] and a sequential-combined approach. In the sequential-combined approach, i3DRefine, GalaxyRefine and ModRefiner are run sequentially, followed by MESHI and FG-MD to fix the secondary structure and remove clash. Each program generates 1000 decoys, resulting in 7000 decoys generated in total.

Clustering and model selection. 7000 decoys are clustered by SPICKER, and the top 5 largest cluster are selected. The five decoys nearest to the five cluster centers (average of all members from the cluster) are picked up as initial models. A short atomic-level molecular dynamics simulation (1ns) by LAMMPS[10] with AMBER99[11] force field is followed to get the final models. The final models are re-ranked by MolProbity score before submitted.

Availability

The FG-MD server is available at http://zhanglab.ccmb.med.umich.edu/FG-MD/.
The ModRefiner server is available at https://zhanglab.ccmb.med.umich.edu/ModRefiner/.

1. Zhang, Y. & Skolnick, J. (2004). SPICKER: A clustering approach to identify near-native protein folds. *Journal of Computational Chemistry* **25**, 865-871.
2. Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S. & Richardson, D. C. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D: Biological Crystallography* **66**, 12-21.
3. Xu, J. & Zhang, Y. (2010). How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **26**, 889-95.
4. Y. Zhang, J. S. (2005). TM-align: A protein structure alignment algorithm based on TM-score. *Nucleic Acids Research* **33**, 2302-2309.
5. Zhang, J., Liang, Y. & Zhang, Y. (2011). Atomic-level protein structure refinement using fragment guided molecular dynamics conformation sampling. *Structure (London, England : 1993)* **19**, 1784-1795.
6. Xu, D. & Zhang, Y. (2011). Improving the Physical Realism and Structural Accuracy of Protein Models by a Two-Step Atomic-Level Energy Minimization. *Biophysical Journal* **101**, 2525-2534.
7. Heo, L., Park, H. & Seok, C. (2013). GalaxyRefine: protein structure refinement driven by side-chain repacking. *Nucleic Acids Research* **41**, W384-W388.
8. Bhattacharya, D., Nowotny, J., Cao, R. & Cheng, J. (2016). 3Drefine: an interactive web server for efficient protein structure refinement. *Nucleic Acids Research* **44**, W406-W409.
9. 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.
10. Plimpton, S. (1995). Fast Parallel Algorithms for Short-Range Molecular Dynamics. *Journal of Computational Physics* **117**, 1-19.
11. Wang, J., Cieplak, P. & Kollman, P. A. (2000). How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *Journal of Computational Chemistry* **21**, 1049-1074.

Protein 3D structure predictions by C-I-TASSER in CASP13

Wei Zheng[1], Chengxin Zhang[1], Yang Li[1,2], S M Golam Mortuza[1], Yang Zhang[1]

1Department of Computational Medicine and Bioinformatics, Department of Biological Chemistry, University of Michigan, 100 Washtenaw Ave, Ann Arbor, MI 48109; 2School of Computer Science and Engineering, Nanjing University of Science and Technology, Xiaolingwei 200, Nanjing, China, 210094

yangzhanglab@umich.edu

The Zhang-Sever structure prediction in CASP13 is based on C-I-TASSER (Zheng et al, in preparation), which is an extended version of the I-TASSER pipeline.[1-3] The query sequence is first threaded by LOMETS[4] through the PDB library to identify putative structure templates. Continuously aligned fragments are excised from the threading templates and used to reassemble full-length models by iterative replica-exchange Monte Carlo (REMC) simulations as described previously.[2, 3, 5] Contact predictions by NeBcon[6] and ResPRE (Li et al, in preparation) are combined and used for guiding the REMC structure assembly simulations. Finally, the simulation decoys are clustered by SPICKER,[7] with the selected models further refined at atomic-level by the fragment-guided molecular dynamic (FG-MD[8]) simulations.

Compared to the I-TASSER pipeline used in CASP12,[9] the major new development in C-I-TASSER is the incorporation of the new contact-map prediction from the ResPRE program, which creates contact-map predictions by combining precision matrix with deep residual neural network training. The posteriors probabilities of ResPRE and four other contact prediction programs (DeepCov,[10] DNCON2,[11] MetaPSICOV2,[12] and GREMLIN[13]) are used by NeBcon as input features to create new composite contact-maps through neural network training.[6] These contacts are incorporated into the C-I-TASSER simulations through a 3-gradient contact potential to guide the replica-exchange Monte Carlo fragment assembly simulations (see QUARK Abstract). Here, the weighting parameters and the number of contacts used in different categories (short-, medium- and long-range) for different programs (ResPRE, NeBcon) are dependent on the length of the query sequence and the confidence score of different programs, which were pre-trained through a non-redundant set of 187 proteins.

In addition to the sequence-based contact-map prediction, the previously-used rectangle-shape potential in the threading-based contact restraints and the generic side-chain and Cα contact potentials are also converted to the 3-gradient contact potential (Zheng et al, in preparation). Accordingly, all the weighting parameters associated with the contact potentials have been systematically retrained using the structural decoys by the maximization of correlation between TM-score and total energy function[14].

1. Zhang, Y., I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 2008, 9, 40.
2. Roy, A.; Kucukural, A.; Zhang, Y., I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 2010, 5, 725-38.
3. Yang, J.; Yan, R.; Roy, A.; Xu, D.; Poisson, J.; Zhang, Y., The I-TASSER Suite: protein structure and function prediction. Nature Methods 2015, 12, 7-8.
4. Wu, S.; Zhang, Y., LOMETS: A local meta-threading-server for protein structure prediction. Nucl. Acids. Res. 2007, 35, 3375-3382.
5. Wu, S.; Skolnick, J.; Zhang, Y., Ab initio modeling of small proteins by iterative TASSER simulations. BMC Biol 2007, 5, 17.
6. He, B.; Mortuza, S. M.; Wang, Y.; Shen, H. B.; Zhang, Y., NeBcon: protein contact map prediction using neural network training coupled with naive Bayes classifiers. Bioinformatics 2017, 33, 2296-2306.
7. Zhang, Y.; Skolnick, J., SPICKER: A clustering approach to identify near-native protein folds. J Comput Chem 2004, 25, 865-71.

8. Zhang, J.; Liang, Y.; Zhang, Y., Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling. Structure 2011, 19, 1784-95.
9. Zhang, C.; Mortuza, S. M.; He, B.; Wang, Y.; Zhang, Y., Template-based and free modeling of I-TASSER and QUARK pipelines using predicted contact maps in CASP12. Proteins 2018, 86 Suppl 1, 136-151.
10. Jones, D. T.; Kandathil, S. M., High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. Bioinformatics 2018, bty341-bty341.
11. Adhikari, B.; Hou, J.; Cheng, J., DNCON2: improved protein contact prediction using two-level deep convolutional neural networks. Bioinformatics 2017.
12. Buchan, D. W. A.; Jones, D. T., Improved protein contact predictions with the MetaPSICOV2 server in CASP12. Proteins-Structure Function and Bioinformatics 2018, 86, 78-83.
13. Kamisetty, H.; Ovchinnikov, S.; Baker, D., Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. Proceedings of the National Academy of Sciences of the United States of America 2013, 110, 15674-15679.
14. Zhang, Y.; Kolinski, A.; Skolnick, J., TOUCHSTONE II: A new approach to ab initio protein structure prediction. Biophys. J. 2003, 85, 1145-1164.

Protein Contact Map Prediction by Coupling Residual Two-Dimensional Neural Networks

J. Hanson[1], K. K. Paliwal[1], T. Litfin[2], Y. Yang[3], Y. Zhou[2]

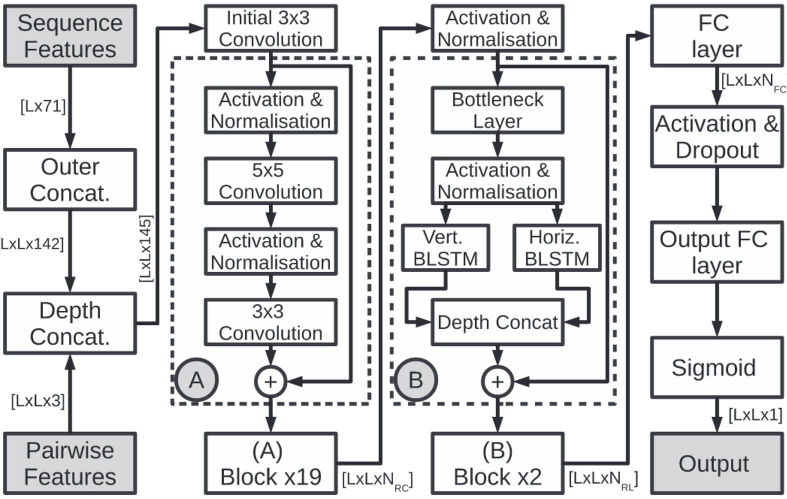
[1]- Signal Processing Laboratory, Griffith University, Nathan, QLD, Australia, [2]- Institute for Glycomics, Griffith University, Southport, QLD, Australia, [3]- School of Data and Computer Science, Sun Yat-Sen University, Guangzhou, Guangdong, China

yaoqi.zhou@griffith.edu.au

Deep learning saw a vast increase in the performance of protein contact map prediction in the CASP12 competition. This was due to their ability to learn very robust and complex relationships between a set of input features and their interactions in forming the folded contact map, especially for proteins with few homologs, where Evolutionary Coupling Analysis (ECA) falls short. The key to continuing this performance rise is in capturing as wide a context for a target residue pair as possible, because structurally but not sequentially neighboring amino acid residues provide key interactions in the folding process. To achieve this, we proposed the coupling of both residual Convolutional Neural Networks (ResNets[1]) and Residual two-dimensional Bidirectional Long Short-Term Memory Networks (2D-BRLSTM’s[2,3]). The intuition behind this methodology is that the ResNet sections will act as motif aggregators due to their ability to easily identify short-term dependencies, while the the 2D-BRLSTM sections will act as propagators of these learned motifs due to their ability to learn sequence-wide dependencies.

Methods

Our method[4] employs an ensemble of 2D-BRLSTM’s coupled with ResNets. Figure 1 shows the architecture for one of the ensemble models for a protein of length *L*, illustrating how we have connected 2D-BLSTM (block B) and ResNet block A) models to capitalize on their individual strengths. Due to their smaller weight space, ResNets are very effective at finding shorter sequential motifs present in the data, whereas recurrent architectures such as BLSTM’s are able to propagate these learned features much deeper throughout the sequence. In this model we adapt the BLSTM architecture to be two-dimensional by employing a horizontal (standard) BLSTM along with a vertical (transposed) BLSTM to capture long-range pairwise dependencies in the contact map.



Our model[4] utilizes the outputs from several programs. The one-dimensional features include the evolutionary profiles from PSI-Blast[5] and HHBlits[6], the predicted local structure from SPIDER3[7], and seven physicochemical properties per residue[8]. We also used the two-dimensional, or pair-wise features provided by the ECA-based contact map predictor CCMpred[9], and the mutual and direct coupling information from DCA[10]. The one-dimensional features are transformed to two-dimensional space by the use of outer concatenation.

Results

Testing on a set of proteins deposited after 2015, we found that this model provided consistently higher precisions at all L/k groupings and sequence position separations. Most importantly, we improved on the already outstanding long-range precisions provided by RaptorX-Contact, the top-performing model in CASP12[11]. We also calculate the area under the receiver operating characteristic (AUC) and precision-recall (AUCPR), curves and find a substantial improvement over all other methods (0.958 to the second best 0.909 for AUC), indicating that the improvement is not biased to positive predictions but based on an overall improvement for the whole contact map.

Availability

The model, called SPOT-Contact, is available as both a webserver and a downloadable standalone at sparks-lab.org, along with the training and testing datasets.

1. He,K.et al. (2016a) Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Las Vegas, USA, pp. 770–778.
2. Schuster,M. and Paliwal,K.K. (1997) Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.*, **45**, 2673–2681.
3. Visin,F.et al. (2015) ReNet: a recurrent neural network based alternative to convolutional networks. *CoRR*, Abs/1505.00393
4. Hanson,J. et al. (2018) Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks., *Bioinformatics*, doi: 10.1093/bioinformatics/bty481
5. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
6. Remmert,M. et al. (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173–175.
7. Heffernan,R. et al. (2017) Capturing non-local interactions by long short term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure. *Bioinformatics*, **33**, 2842–2849.
8. Meiler,J. et al. (2001) Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. *Mol. Model. Annu.*, **7**, 360–369.
9. Seemayer,S. et al. (2014) CCMpredfast and precise prediction of protein residue–residue contacts from correlated mutations. *Bioinformatics*, **30**, 3128–3130.
10. Morcos,F. et al. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci.*, **108**, E1293–E1301.
11. Wang,S. et al. (2018) Analysis of deep learning methods for blind protein contact prediction in casp12. *Prot. Struct. Funct. Bioinform.*,**86**, 67–77.

CAMEO - Continuous Automated Model EvaluatiOn

Juergen Haas[1,2], Rafal Gumienny[1,2], Dario Behringer[1,2], Torsten Schwede[1,2]

[1-] SIB Swiss Institute of Bioinformatics, [2] Biozentrum, University of Basel

Juergen.Haas@unibas.ch

CAMEO - Continuous Automated Model EvaluatiOn (http://www.cameo3d.org) – continuously assesses the prediction performance of servers participating in one of the three categories: 3D - Protein Structure Prediction, QE – Quality Estimation and CP – Contact Prediction. CAMEO[1] is based on the weekly pre-release of amino acid sequences and ligand identity for experimental protein structures which are going to be part of the subsequent PDB release in the following week. Each Saturday, about 20 protein sequences are selected as the target set and are then submitted to the participating servers for CAMEO 3D and CAMEO CP categories and predictions are collected during the next four days. For CAMEO QE models of the public servers registered with CAMEO 3D are submitted 12h after the sequences have been submitted to CAMEO 3D. For the evaluation of the predictions, the coordinates released by the PDB on Wednesday are used as reference, and the assessments are then published on cameo3d.org for follow-up performance analyses.

Since CAMEO is linked to the PDB release cycle, a large number of targets have been evaluated. For CAMEO 3D 6271 targets were evaluated during the last 350 weeks. The target set is diverse - consisting of 1,762 hard (lDDT*<50), 3224 medium and 1225 easy targets (lDDT >=75). Overall, 2607 targets were homo-oligomers - allowing to assess the ability of servers to correctly predict the quaternary state of a target protein.

The assessment of residue-residue contact predictions in CAMEO CP covered 330 targets collected over 53 weeks. The latest category reflects recent findings, that the quality of a 3D model can be improved greatly by considering predicted residue-residue contacts during the modeling process [3]. This applies in particular for target proteins larger than 250 amino acid residues, with little structural templates coverage.

Within the ELIXIR framework CAMEO is currently being integrated into OpenEBench (OEB)[4] – the infra-structure to take benchmarking to the next level. It supports all life science communities, includes technical and scientific aspects and is envisioned to cover three levels of complexity. Level 1 translates to sharing of existing assessment data with the benefit of a unified interface and metrics that are available across communities for straightforward identification of quality. Technically, it is based on a generalized data model that captures the entire benchmarking procedure. Level 2 is implemented as a current prototype, where within one community scientists can bring their own data and OEB allows to run the community-supported assessment workflows, thereby integrating the new data with existing data sets on OEB. The Level 3 concept is currently in planning phase. It would run assessment workflows regularly within OEB alleviating the tedious setup of a benchmarking event, which is especially interesting for long-term commitments, as the community can benefit of standardized processes.

The aim of CAMEO is to support various aspects of protein structure modeling by providing a spectrum of different scores. CAMEO supports the developers of prediction servers by rapidly assessing new developments anonymously and monitoring the performance of their public productive servers continuously. CAMEO allows life scientists to better understand which public modeling server is the most suited for their specific use case. CAMEO stimulates the respective prediction communities in discussing new scores, thereby covering yet another aspect of the respective field. We hence invite the community to discuss scoring schemes and emerging methods to best reflect recent scientific developments.

*average local Difference Distance Test[2] across all models for a given target.

1. Haas,J. et al. (2018) Continuous automated model evaluation (CAMEO) complementing the critical assessment of structure prediction in CASP12. *Proteins* **86**(Suppl. 1), 387–398.
2. Mariani,V. et al. (2013) lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **29**, 2722–2728.
3. Monastyrskyy B. et.al. (2016) New encouraging developments in contact prediction: Assessment of the CASP11 results. *Proteins* **84**(Suppl 1):131–144.
4. Capella S, et.al. (2017) Lessons Learned: Recommendations for Establishing Critical Periodic Scientific Benchmarking. *bioRxiv* 181677.

ProMod3 and OpenStructure – the SWISS-MODEL toolbox to generate, manipulate and compare protein models

G. Tauriello[1,2], G. Studer[1,2], M. Bertoni[1,2], R. Gumienny[1,2], S. Bienert[1,2], T. Schwede[1,2]

1- Biozentrum, University of Basel, 2- SIB, Swiss Institute of Bioinformatics

gerardo.tauriello@unibas.ch

Since 25 years, SWISS-MODEL[1] has made protein modeling accessible to all life science researchers worldwide. Over the past 12 months, we generated approximately 1.26 million models, i.e. 2.4 models per minute have been requested by our users through our interactive modeling services. We furthermore provide up-to-date models for more than 220'000 protein sequences of 12 selected core species in our SWISS-MODEL Repository[2]. All models can be used for any purpose, even commercially, as they are licensed under the CC BY-SA 4.0 license. Our focus lies on high-quality template-based modeling with a short response time and on modeling of macromolecular complexes.

To be able to provide the SWISS-MODEL service, we rely on in-house developed software. The software needs to be flexible enough to accommodate our development needs, but also stable, reliable and fast for "production use" in SWISS-MODEL. Here, we present the first public release of our new modeling engine ProMod3[3] and the latest developments in the OpenStructure computational structural biology framework[4].

Methods

ProMod3 is designed as an extension to OpenStructure and has been the modeling engine of SWISS-MODEL since June 2016. Given a target sequence, a template structure and an alignment between them, it reliably produces a gap-free, all-atom model for the target protein. The modeling engine performs the following steps: (1) building an initial model by transferring conserved parts from the template structure, (2) loop modeling to resolve insertions and deletions using fragments in a structural database, (3) modeling of side chains using the backbone dependent rotamer library from the Dunbrack group[5] and minimizing the SCWRL4 energy function[6], (4) energy minimization using the CHARMM27 force field[7].

OpenStructure now includes a command-line tool to compare structures using our superposition-free, all-atom lDDT-[8] and QS-scores[9]. In the process, we also introduced new lDDT-scores which are suited to compare macromolecular complexes. We furthermore included wrappers around the OpenMM molecular mechanics library[10] to perform energy minimization tasks to resolve stereochemical irregularities and clashes introduced in the modeling process.

To ensure efficiency, most algorithms and data structures of our software tools are implemented in C++ and made available to the Python scripting language which enables fast prototyping.

Results

To benchmark the performance of ProMod3, 226 target sequences submitted by CAMEO[11] have been selected. The best target-template-alignment according to the e-value from HHblits[12] was used to produce one model per target with MODELLER[13] and with ProMod3 using the exact same input data. Modeling accuracy has been measured by the lDDT-score. The MolProbity overall score has been used to evaluate stereochemistry as an additional but equally important aspect. ProMod3 shows an improvement in both metrics: the average increase in lDDT score is 1.68 and the average decrease in MolProbity score is 1.27. Besides producing better results, ProMod3 is also faster by a factor of 1.4x.

ProMod3 is furthermore continuously evaluated as part of SWISS-MODEL within the CAMEO project. In CAMEO results, we observe that SWISS-MODEL has the lowest response time to generate models and excels at model quality for binding sites (IDDT-BS), for high-quality models (lDDT for "Easy" targets) and for quaternary structure predictions (QS-score).

Availability

To enable other modeling developers to use our software easily, we provide Docker and Singularity images. In future updates, we will also include our model quality estimation tools (see abstract by Studer et al.). SWISS-MODEL is available as a web service at https://swissmodel.expasy.org.

1. Waterhouse,A. et al. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res*. **46**(W1), W296-W303 (2018).
2. Bienert,S. et al. (2017). The SWISS-MODEL Repository - new features and functionality. *Nucleic Acids Res*. **45**, D313-D319.
3. Studer,G. et al. ProMod3 - A Versatile Homology Modelling Toolbox. In preparation.
4. Biasini,M. et al. (2013). OpenStructure: an integrated software framework for computational structural biology. *Acta Cryst*. **69**, 701-709.
5. Shapovalov,M.V. et al. (2011). A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure* **19**, 844–858.
6. Krivov,G.G. et al. (2009). Improved prediction of protein side-chain conformations with SCWRL4. *Proteins* **77**, 778–795.
7. Mackerell,A.D. et al. (2004). Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comput. Chem.* **25**, 1400–1415.
8. Mariani,V. et al. (2013). lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **29**, 2722–2728.
9. Bertoni,M. et al. (2017). Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. *Scientific Rep.* **7**, 10480.
10. Eastman,P. et al. (2017). OpenMM 7: rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* **13**, e1005659.
11. Haas,J. et al. (2018) Continuous automated model evaluation (CAMEO) complementing the critical assessment of structure prediction in CASP12. *Proteins* **86**(Suppl. 1), 387–398.
12. Remmert,M. et al. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* **9**, 173-175.
13. Sali,A. et al. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815.