A Contact-Assisted Approach to Protein Structure Structure Prediction and Its Assessment in CASP10

Badri Adhikari, Xin Deng, Jilong Li, Debswapna Bhattacharya, and Jianlin Cheng*
Department of Computer Science, University of Missouri, Columbia, MO 65211 USA
*Corresponding author: chengji@missouri.edu

Abstract

Among different approaches to predict the 3D structure of a protein, one important idea is to predict a protein residue-residue contact map and then construct a full 3D structure from the contact-map. Instead of building a structure purely from contacts information, here we describe a contact-assisted structure prediction approach that uses only a few known contacts to improve the quality of already predicted models. Our approach for contact assisted structure prediction uses a novel method for selecting and refining protein structural models. With input test data as the predicted structures for 15 protein targets used in the contact-assisted prediction category in the 10th Critical Assessment of Techniques for Protein Structure Prediction (CASP10), we demonstrate that weighted contacts satisfaction score along with other established model quality assessment scores is a promising technique for selecting good structures and ultimately for better structure prediction.

Availability:
http://protein.rnet.missouri.edu/contact_assisted/index.html

Introduction

The problem of predicting 3D protein structure from amino acid sequence is currently a great challenge in structural bioinformatics. Among popular methods to predict a protein's structure, are the methods that use residue-residue contact maps. A contact map of a 3D structure of a protein is a binary two dimensional matrix M where M[i,j] is 1 or 0, based on whether or not the Euclidean distance between the residues i and j in the Cartesian space is less than or equal to a predefined distance threshold (e.g. 8 Angstrom). The idea of using contact-map to solve the problem of protein folding as was introduced back in 1971 (Nishikawa et al. 1972) and is still actively being explored. The principle behind these contact-based methods is to predict a contact map and then construct a full 3D structure from this contact map.

Although the accuracy of contact map prediction is generally too low to be used as the only source of information to accurately construct a protein structure in most cases, some interesting results of constructing 3D structures from contact maps have been observed (Vassura et al. 2008). The technique, which first predicts more accurate residue-residue contacts for some proteins with a large family of known sequences that contains rich evolutionary information and then predicts the full structure from contacts along with other information, has recently been exploited to predict 3D protein structure with root-mean square deviation (RMSD) from experimental structure of about 2.7 Å (Marks et al. 2011). Instead of using the whole contact-map, a small portion of useful contacts can also be effectively used in the structure prediction process as demonstrated by (Skolnick et al. 1997). For example, using just 20 restraints, myoglobin (146 residue long helical protein) can be folded to structures whose average RMSD from experimental structures is 5.65 Å (Skolnick et al. 1997).

The idea of using only a relatively small number of contacts as additional information to aid protein structure prediction is gaining more interest since the recent introduction of contact assisted protein structure prediction in the 10th Critical Assessment of Techniques for Protein Structure Prediction (CASP10) in 2012. The contact-assisted structure modeling experiment in CASP10 was designed to test how the knowledge of several long-range contacts influences the ability of predictors to model a complete protein structure. The category had total 15 target chains consisting of 17 domains and most of the targets were hard targets as they were either free-modeling targets or hard template-based modeling targets. For each target, 3 to 34 known contacts were given to aid tertiary structure prediction, except for target Tc653, which instead of contacts, had no-contacts information (i.e. list of pair of residues that are not in contact in the native structure). Before a target was released along with some contacts in the contact-assisted category, the same target had already been released as a normal tertiary structure prediction target (e.g. template-based modeling or template-free modeling). This let the CASP assessors to check the improvement in model quality between the models predicted with some known contacts and that those without any contact information. In this paper, we describe our approach implemented for contact-assisted protein structure prediction for these CASP 10 targets, and discuss further prospects of the method.

Methods

Overview

Our method uses known residue-residue contacts together with a pool of pre-constructed structure models as input to predict protein structures for a target. During the CASP10 experiment, before a target was released in the contact-assisted category along with contacts, the same target had been released as regular target and the structural models for the target predicted by the tertiary structure predictors participating in CASP10 were publicly accessible at CASP10 website. On average each target had about 250 predicted models, which were in a wide range of quality whose GDT-TS score range from 0.006 to 0.763. These models were used as the input set of structural models for our contact-assisted prediction method.

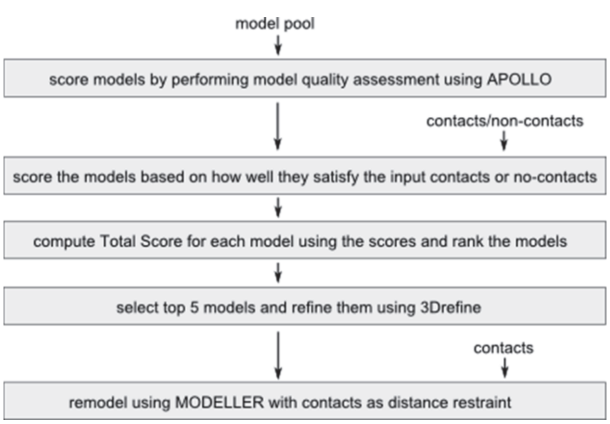

Figure 1 The five steps of our contacts assisted structure prediction method.

As shown in Figure 1, our method for contact assisted protein structure prediction is comprised of 5 steps: (1) perform model quality assessment using APOLLO (Wang et al. 2011) to assess the input set of models and score them; (2) score the models based on how well they satisfy the input contacts or no-contacts; (3) rank the models by integrating the scores obtained during the previous two steps (i.e., Apollo's GDT-TS score (Zemla et al. 1999), Apollo's MaxSub score (Siew et al. 2000), Apollo's TM-score (Zhang and Skolnick 2004), percent of exact contacts satisfied, percent of no-contacts satisfied); (4) select top 5 models and refine them using 3Drefine (Bhattacharya and Cheng 2013) ; and (5) remodel the top 5 models using Modeller with contacts as distance restraint. The first three steps form the model selection process and the last two steps the model refinement process, which are described in more details in the two sub-sections that follow.

Contact-Assisted Model Selection

The programs used for model quality assessment or computing the accuracy of models in a model pool without knowing the native structure, commonly known as Model Quality Assessment Programs (MQAPs) (Kihara et al. 2009), predict either global quality of the entire model, or residue-specific local qualities, or even both. We used APOLLO (Wang et al. 2011) as a component of our model selection method. When a pool of models is supplied as input to APOLLO, it performs a full pair-wise comparison between all the models using TM-Score tool (Zhang and Skolnick 2004) and assigns scores including average pair-wise GDT-TS score, average pairwise TM-Score (Zhang and Skolnick 2004), and average MaxSub score to each model. Models similar to most of the other models get higher evaluation scores. For example, the final TM-score for a model in a model pool is the average of all TM-scores between this model and all other models.

$$Total\ Score = \underbrace{GDT\text{-}TS\ score + MaxSub\ score + TM\text{-}score}_{\text{APOLLO component}} + \underbrace{(\%\ of\ contacts\ satisfied) - 0.1 * (\%\ of\ no\text{-}contacts\ satisfied)}_{\text{Contacts component}} \quad (1)$$

In addition to the 3 scores from APOLLO, we used two contact terms to account for the contacts / no-contacts that a model satisfies. The first contact term, percent of contacts satisfied, is the number of known contacts present in the model divided by total number of given contacts. The second contact term is a negative score equal to the percent of known no-contacts realized as contacts in the model. The two contact / no-contact terms and the three APOLLO's terms were summed into a total score to rank input models of a target according to the formula in Equation (1). The weights for the terms in the equation were chosen empirically based on our experience.

Contact Assisted Structure Modeling

The top five selected models are first refined by 3Drefine in two steps: optimization of hydrogen bond network followed by atomic-level energy minimization using physics-based and knowledge-based energies (Bhattacharya and Cheng 2013). To refold the refined models, contacts are transformed to distance restrains, and then structure modeling program, MODELLER (Eswar et al. 2007), a program for homology or comparative modeling of protein three-dimensional structures, is used. MODELLER implements comparative protein structure modeling by satisfaction of spatial restraints and additional distance restraints can be incorporated easily. Typically, to make a structure prediction, MODELLER requires structure templates along with an alignment file that contains the alignment of the input sequence aligned with the sequences of the template structures. For each prediction, we used the selected model as the only template structure and then created an alignment file that has the input sequences aligned fully with the template sequences (e.g. themselves). The default “automodel” modeling protocol in MODELLER was used with additional distance restraints derived from provided contacts. A residue-residue contact was converted to 8.0 angstrom mean distance between Cß-Cß atoms (or Ca atom in case of GLY residue). The standard deviation of the distance is set to 0.1 Angstrom and a harmonic potential function was applied to enforce the distance restraint. In this way, a refined model was refolded using MODELLER, except for target Tc653, which had only no-contacts as input.

Results and Discussions

The CASP10 targets (either full proteins or domains) and the corresponding contact information used to benchmark our contact-assisted protein structure prediction method are listed in Table 1.

In order to evaluate how well our contact-assisted model selection method ranked the models, for each target, we ranked its input models based on their real GDT-TS score obtained by comparing them with the native structures, and marked the top 1 model picked by our scoring function (see Figure 2). In addition, to check how well the components of the scoring function ranked the models, we also marked the models picked by these components separately. The average correlation between the actual GDT-TS scores and the predicted total scores was 0.601 as shown in Table 2.

Despite the improvement in average correlation, the contribution of the contact component in ranking models was not consistent. In some case, it ranked a good-quality model at the top, but in another case, it may select a low-quality model at the top (loss is shown in Table 2). Thus, how to more effectively use known contacts with other model quality assessment methods in model ranking is still an issue yet to solve.

During the CASP10 prediction season, the last step of our method (remodeling with MODELLER) was being developed as the CASP experiment was proceeding, and was only ready for being applied to the last two targets. Thus, we applied the fully developed method to the missed targets to generate the final models again after the CASP10 was over in order to evaluate our method.

Target #	**Target**	**# of residues**	**# of contacts / no contacts**
1	T0649	184	16
2	T0653	383	12
3	Tc658-D1	166	16
4	Tc666	180	14
5	Tc673	62	5
6	Tc676	173	17
7	Tc678	154	12
8	Tc680	96	3
9	Tc684-D1	73	8
9	Tc684-D2	168	18
10	T0691	141	15
11	Tc705-D2	344	34
12	Tc717-D2	166	15
13	Tc719-D6	163	13
14	Tc734	212	20
15	Tc735-D1	233	28
15	Tc735-D2	88	7

Table 1 Targets in contact-assisted structure modeling category in CASP10. For target Tc653 no contacts were provided instead of contacts (Source: http://predictioncenter.org/casp10/doc/presentations/CASP10_contact_assisted_BKL.pdf)

Ranking Method	**Average Correlation**	**Average Loss**
Total Score Formula	0.601	0.088
APOLLO Component only	0.559	0.087
Contacts Component only	0.390	0.088

Table 2 Average correlation column is the Pearson Correlation between actual GDT-TS scores and GDT-TS scores ranked by the ranking method used. Loss column is the difference between the GDT-TS score of the best model and the top 1 model ranked by the method used.

	RMSD	GDT-TS	GDT-HA
Model ranked top 1 using Total Score	17.33	0.3079	0.1738
After refinement	17.33	0.3090	0.1717
After remodeling using MODELLER	7.789	0.3498	0.1835

Table 3 Stepwise evaluation of the prediction of first domain (D1) of Target Tc735. In this example, significant improvement is observed in the model quality after remodeling.

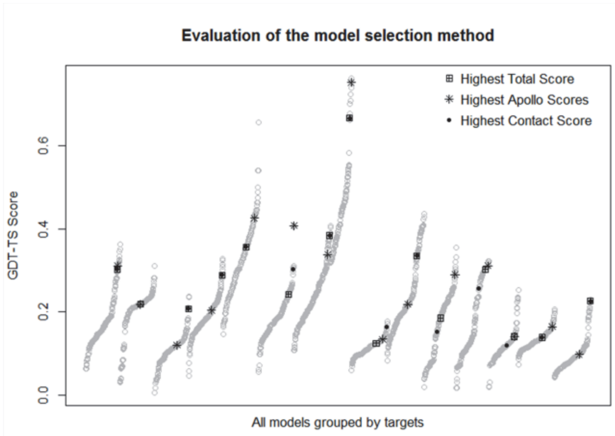

Figure 2. Evaluation of the scoring function of ranking models. Y-axis denotes real GDT-TS scores and X-axis indices of the models. Each group of models represents the models for a target, ordered according to their real GDT-TS scores. In each group, the top models selected by the total score, the APOLLO component, and the contact components were marked by three legends, respectively. The second group of models does not have a highest contact scoring model because only no-contacts were provided for this target.

To observe the stepwise improvement in the models, we compared our refined models (generated in step 4) and the re-folded models (generated in step 5) with the native structures. As shown in Table 4, the refinement step with 3DRefine slightly improves the quality of the selected models, most of the times, with the average RMSD improvement of 0.0035. The final step of using MODELLER mostly improves the quality of the model, with average RMSD improvement of 2.3027, and sometimes the improvement was drastic in terms of RMSD.

Despite not being conclusive, the results seem to show that using contacts as an additional measure to refold models with existing modeling techniques such as MODELLER can be a promising approach to embedding a few known contacts into existing protein structure prediction methods to improve the overall prediction accuracy. Here, target Tc735 is analyzed as a case to study the potential effectiveness of the contact-assisted prediction method. The target Tc735 has two structural domains: D1 and D2. We consider the first domain of this target (residues 29 - 262) as an example to demonstrate the application of our method. As shown in Table 3, 10+% improvement is observed in the model quality as the model was improved from GDT-TS score of 0.3079 to 0.3498.

#	Target	Selected Model		Refined Model		Modeller Model		Final Improvement	
		RMSD	GDT-TS	RMSD	GDT-TS	RMSD	GDT-TS	Change in RMSD	Change in GDT-TS
1	Tc649	13.880	0.3026	***13.790***	***0.3026***	10.370	0.2632	3.510	-0.039
2	Tc653	18.780	0.2190	***18.770***	***0.2203***	-	-	-	-
3	Tc658	14.220	0.2078	***14.220***	***0.2078***	13.060	0.2078	1.160	0.000
4	Tc666	8.355	0.2889	***8.366***	***0.2917***	8.005	0.3056	0.350	0.017
5	Tc673	9.384	0.3566	***9.399***	***0.3525***	9.249	0.3566	0.135	0.000
6	Tc676	11.480	0.2425	***11.470***	***0.2455***	11.110	0.2470	0.370	0.005
7	Tc678	6.980	0.3842	***6.966***	***0.3826***	6.719	0.3960	0.261	0.012
8	Tc680	2.979	0.7473	***2.956***	***0.7446***	4.132	0.6979	-1.153	-0.049
9	Tc684	20.460	0.1250	***20.460***	***0.1239***	18.640	0.1164	1.820	-0.009
10	Tc691	9.619	0.3358	***9.597***	***0.3376***	9.514	0.3431	0.105	0.007
11	Tc705	13.440	0.1853	13.450	0.1875	11.020	0.1969	2.420	0.012
12	Tc717	18.050	0.1747	18.070	0.1747	16.440	0.1687	1.610	-0.006
13	Tc734	17.620	0.1392	17.660	0.1403	16.290	0.1439	1.330	0.005
14	Tc719	23.140	0.1411	23.170	0.1380	***15.410***	***0.1212***	7.730	-0.020
15	Tc735	25.150	0.2269	25.140	0.2269	***12.560***	***0.2580***	12.590	0.031
						Average Change:		2.302	-0.002

Table 4 Evaluation of the top 1 prediction for all targets. Selected Models column shows the RMSD and GDT-TS score of the top 1 ranked model, selected by our Total Score formula, compared with the native structure. Refined Models column shows the RMSD and GDT-TS score of the top 1 ranked model after refinement. Final Improvement column shows the improvement in RMSD and GDT-TS after remodeling with MODELLER. Highlighted models are the models sent to the CASP10 competition. Re-modeling was not performed for targets Tc653 because no contacts were provided for this target. Targets Tc705, Tc717 and Tc734 were missed by mistake during the CASP10 experiment and so were not sent to CASP10.

The RMSD of the model was reduced from 17.33 Angstrom to 7.79 Angstrom. Figure 3 shows that remodeling appears to bring some poorly modeled regions (e.g. one terminal region is folded inside) closer to the native structure.

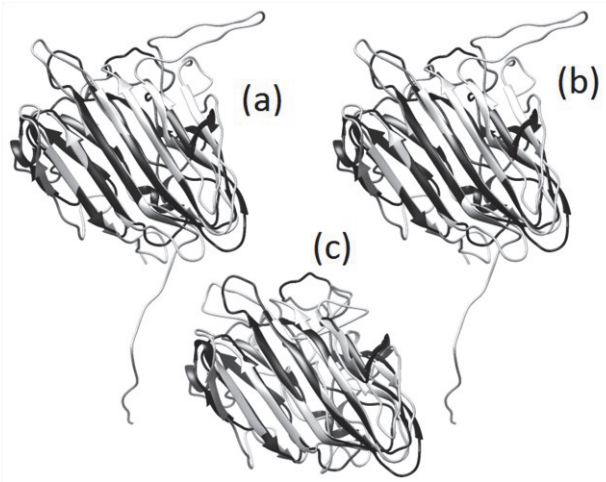

Figure 3 Prediction of first domain of target Tc735 using our method. (a) model ranked top 1 by our Total Score formula in orange superimposed with native structure in dark (b) Same structure after refinement in red superimposed with native in dark (c) Same structure re-folded using MODELLER with contacts as distance restraints superimposed with native in dark.

Conclusion and Future Works

In this work, we report a simple structure prediction method using a small number of known residue-reside contacts / no-contacts to aid protein structure prediction. The preliminary results (with challenging benchmark data comprised of hard targets) demonstrated that the known contacts could be incorporated into existing protein structure prediction techniques to improve protein model ranking and generation in some situations, suggesting contact-assisted protein structure prediction may be a promising technique to enhance protein structure modeling. We expect that more advanced methods can be developed to better use contact information to more substantially improve protein structure prediction.

We are currently working on using predicted contacts to guide ab initio protein structure prediction. Contacts based energy function will be optimized using simulated annealing with energy minimization techniques like Powell minimization (Powell 1964) as demonstrated in (Marks et al. 2011) and/or limited memory BFGS minimization (Nocedal 1989). We also plan to experiment the combination of fragment replacement approach for ab initio structure prediction and optimization using contact based energy function, and structural profile based conformation sampling as discussed in (Olson et al. 2012). Through the optimization process, we aim to satisfy as many supplied contacts as possible. In addition to the contacts as guiding energy function, we plan to use additional information like predicted secondary structure in the form of distance and angular restraints to improve the quality of the models. The completed work described above and the work in progress together will demonstrate approaches of using contacts information to predict protein 3D structures in the field of template-based modeling as well as template-free modeling.

Acknowledgement

The work was partially supported by a NIH NIGMS grant (R01GM093123) to JC.

References

Bhattacharya, D. and Cheng, J. 2013. 3Drefine: consistent protein structure refinement by optimizing hydrogen bonding network and atomic-level energy minimization., *Proteins*, 81(1), 119–31, doi:10.1002/prot.24167.

Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M., Pieper, U. and Sali, A. 2007. Comparative protein structure modeling using Modeller, *Current Protocols in Protein Science*, 2.9. 1–2.9. 31.

Kihara, D., Chen, H. and Yang, Y. D. 2009. Quality assessment of protein structure models, *Current Protein and Peptide Science*, 10(3), 216–228 [online] Available from: http://www.ncbi.nlm.nih.gov/pubmed/19519452.

Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. a, Pagnani, A., Zecchina, R. and Sander, C. 2011. Protein 3D structure computed from evolutionary sequence variation, *PloS one*, 6(12), e28766, doi:10.1371/journal.pone.0028766.

Nishikawa, K., Ooi, T., Isogai, Y. and Saitô, N. 1972. Tertiary structure of proteins. I. Representation and computation of the conformations, *Journal of the Physical Society of Japan*, 32(5), 1331–1337.

Nocedal, D. C. L. J. 1989. ON THE LIMITED MEMORY BFGS METHOD FOR LAR GE SCALE OPTIMIZA TION,,.

Olson, B., Molloy, K., Hendi, S. F., & Shehu, A. 2012. Guiding probabilistic search of the protein conformational space with structural profiles. *Journal of bioinformatics and computational biology*, 10(3), 1242005.

Powell, M. J. D. 1964. An efficient method for finding the minimum of a function of several variables without calculating derivatives, *The Computer Journal*, 7(2), 155–162, doi: 10.1093/comjnl/7.2.155.

Siew, N., Elofsson, A., Rychlewski, L. and Fischer, D. 2000. MaxSub: an automated measure for the assessment of protein structure prediction quality, *Bioinformatics*, 16(9), 776–785, doi: 10.1093/bioinformatics/16.9.776.

Skolnick, J., Kolinski, A. and Ortiz, A. R. 1997. MONSSTER: a method for folding globular proteins with a small number of distance restraints, *Journal of molecular biology*, 265(2), 217–241, doi: 10.1006/jmbi.1996.0720.

Vassura, M., Margara, L., Di Lena, P., Medri, F., Fariselli, P. and Casadio, R. 2008. Reconstruction of 3D structures from protein contact maps., *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, 5(3), 357–67, doi: 10.1109/TCBB.2008.27.

Wang, Z., Eickholt, J. and Cheng, J. 2001. APOLLO: a quality assessment service for single and multiple protein models, *Bioinformatics*, 27(12), 1715–1716, doi: 10.1093/bioinformatics/btr268.

Zemla, A., Venclovas, C., Moult, J. and Fidelis, K. 1999. Processing and analysis of CASP3 protein structure predictions., *Proteins*, Suppl 3, 22–9 [online] Available from: http://www.ncbi.nlm.nih.gov/pubmed/10526349 (Accessed 26 March 2013).

Zhang, Y. and Skolnick, J. 2004. Scoring function for automated assessment of protein structure template quality., *Proteins*, 57(4), 702–10, doi:10.1002/prot.20264.