Project Title:

Protein Folding Prediction Using X-ray Diffraction Data as Constraints

Name:○Kam Zhang, Rojan Shrestha, Francois Berenger, David Simoncini, Taeho Jo
Laboratory:
Zhang Initiative Research Unit, RIKEN Advanced Science Institute, RIKEN Wako Institute

1. Background and purpose of the project, relationship of the project with other projects

The protein-folding problem of how the primary sequence determines its tertiary structure is one of the great challenges in computational biology. It is known that all the information required specifying the tertiary structure of a protein is encoded in its primary sequence. Moreover, "structure determines function" is a well-established paradigm. Our ability to predict the structures of proteins from their sequences will greatly facilitate our understanding of the important biological functions that proteins play in the cell.

There are two fundamental challenges in protein structure prediction. One is the construction of precise energy functions that could be used to assess the thermodynamic stability of a protein at a given conformation state. The other is to find the global minimum energy conformation in the complex energy landscape. Both are formidable tasks to tackle. First, the basic physical forces that govern atomic interactions are incompletely and poorly understood. Secondly, it is computationally prohibitive to search for the global minimum energy conformation even if the precise energy function were available.

Recent advancement in computational methods for protein structure prediction has made it possible to generate high quality *de novo* models required for *ab initio* phasing of crystallographic diffraction data using molecular replacement. Despite those encouraging achievements in *ab initio* phasing using *de novo* models, its success is limited only to those targets for which high quality *de novo* models can be generated. In order to increase the scope of targets for which the *ab initio* phasing with *de novo* models can be successfully applied, it is necessary to reduce the errors in the *de novo* models that are used as templates for molecular replacement.

2. Specific usage status of the system and calculation method

Fragment assembly is a powerful method of protein structure prediction that builds protein models from a pool of candidate fragments taken from known structures. Stochastic sampling is subsequently used to refine the models. The structures are first represented as coarse-grained models and then as all-atom models for computational efficiency. Many models have to be generated independently due to the stochastic nature of the sampling methods used to search for the global minimum in a complex energy landscape.

We have developed a method called *EdaFold*, which is a fragment-based approach that shares information between the generated models and steers the search towards native-like regions. A distribution over fragments is estimated from a pool of low energy all-atom models. This iteratively-refined distribution is used to guide the selection of fragments during the building of models for subsequent rounds of structure prediction.

We have also developed an approach called *MORPHEUS* that can identify and rebuild the residues with larger errors, which subsequently reduces the overall C-alpha root mean square deviations (CA-RMSD) to the native protein structure. The error in a predicted model is estimated by the average pairwise geometric distance per residue computed among selected

lowest energy coarse-grained models. This score is subsequently employed to guide a rebuilding process that focuses on more error-prone residues in the coarse-grained models.

3. Result

The *EdaFold* program was evaluated on a benchmark of 20 proteins. Results were compared with results obtained with Rosetta's protocol and showed improved concentration of near-native models. The models generated with both approaches were used to solve the crystallographic phase problem using the Phaser program. In this stringent test, the models generated with our approach obtained a higher success rate.

We have used this *EdaFold* method to participate in the recent "10th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP10)". Our method was ranked No. 1 out of 143 groups from world-wide participants in the template-free modeling category as judged by the average Z-score on GDT_TS. This prospective exercise has further validated the utility of our method.

Our proposed model rebuilding methodology (*MORPHEUS*) has been tested on ten protein targets that were unsuccessful with previous methods. The CA-RMSD of coarse-grained models was improved on average from 4.93Å to 4.06Å. For those models with CA-RMSD less than 3.0Å, their average CA-RMSD was improved from 3.38Å to 2.60Å. These rebuilt coarse-grained models were then turned into all-atom models and refined to produce improved *de novo* models for molecular replacement. Seven diffraction datasets were successfully phased using rebuilt *de novo* models indicating the improved quality of these rebuilt *de novo* models and the effectiveness of this rebuilding process.

4. Conclusion

The use of an estimation of distribution algorithm enabled *EdaFold* to reach lower energy levels and to generate a higher percentage of near-native models. *EdaFold* uses an all-atom energy function and produces models with atomic resolution. We observed an improvement in energy-driven blind selection of models on a benchmark of 20 in comparison with the *Rosetta* AbInitioRelax protocol. Furthermore, *EdaFold* has achieved top ranking in a prospective protein structure prediction competition, CASP10.

We have noticed molecular replacement with *de novo* models were highly sensitive towards the input models. Correct input models were diverged from the target structure mainly because of divergence by local residues. We have found that reducing structural divergence due to local residues significantly improves the quality of global conformations. Usage of those models increases success rate for solving phase problem using molecular replacement.

5. Schedule and prospect for the future

We plan to use the principle of Estimation of Distribution Algorithm for protein design. The problem of protein design can be considered as the inverse problem of protein folding. The sampling algorithm of Estimation of Distribution should also be applicable.

By combining the improvement in main-chain and side-chain sampling, the detection of low accuracy regions in an ensemble of predicted structures and the development of methods to select best decoys either by clustering or using model quality assessment tools, we aim to obtain better protein structure models. These improved structure models will facilitate the molecular replacement solution to the phase problem giving a diffraction dataset.

6. If no job was executed, specify the reason.

N/A.

RICC Usage Report for Fiscal Year 2012

Fiscal Year 2012 List of Publications Resulting from the Use of RICC

[Publication]

Simoncini, D., Berenger, F., Shrestha, R., Zhang, K. Y. J. (2012) A probabilistic fragment-based protein structure prediction algorithm. *PLoS ONE*, 7, e38799, 1-11.

Shrestha, R., Simoncini, D., Zhang, K. Y. J. (2012) Error-estimation-guided rebuilding of *de novo* models increases the success rate for *ab initio* phasing. Acta Cryst. **D68**., 1522-1534.

[Proceedings, etc.]

None.

[Oral presentation at an international symposium]

1. The First BMIRC International Symposium on Frontiers in Computational Systems Biology and Bioengineering, Feb. 28 - Mar. 1, 2013, Fukuoka, Japan. Poster presentation. David Simoncini, Kam Y. J. Zhang, “EdaFold: A Probabilistic Fragment-based Method for Protein Structure Prediction”.
2. Asian Crystallographic Association Meeting, Dec. 2-6, 2012, Adelaide, Australia. Invited Speaker, “Error estimation guided rebuilding of de novo models for ab initio phasing”.
3. 10th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction, Dec. 9-12, Gaeta, Italy, Poster presentation. David Simoncini, Arnout R.D. Voet, Kam Y. J. Zhang, “CASP10 predictions using EdaFold”.
4. Joint Conference on Informatics in Biology, Medicine and Pharmacology, Oct. 14-17, 2012, Tokyo, Japan, Poster presentation, David Simoncini, Francois Berenger, Rojan Shrestha and Kam Y. J. Zhang, “Probabilistic Fragment-based Protein Structure Prediction Algorithm”.
5. European Conference on Computational Biology, Sept. 9-12, 2012, Basel, Switzerland, Poster presentation, Rojan Shrestha, David Simoncini and Kam Y. J. Zhang, “Error estimation guided rebuilding of *de novo* models increases the success rate for *ab initio* phasing”.
6. International Workshop on New Developments of Methods and Software for Protein Crystallography, Aug. 20-26, 2012, Xian, China. Invited Speaker, “A Model Rebuilding Method for *ab initio* Phasing with *de novo* Models”.
7. 20th Annual International Conference on Intelligent Systems for Molecular Biology, July 13-17, Long Beach, California, USA. Invited Speaker, “A Probabilistic Fragment-based Protein Structure Prediction Algorithm”.
8. Pacific Symposium on Biocomputing, Jan. 3-7, 2012, Hawaii, USA, Poster presentation, Simoncini D., Berenger F. C., Shrestha R., Jo T., and Zhang K.Y.J., “A probabilistic fragment-based protein structure prediction algorithm”.