Project Title:

Protein Folding Prediction Using X-ray Diffraction Data as Constraints

Name: ○**Kam Zhang, Rojan Shrestha, Francois Berenger, David Simoncini**

Laboratory at RIKEN: **Zhang Research Initiative Unit**

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. Our ability to predict the structures of proteins from their sequences will enable us to understand of the 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*$_{AA}$, 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 a method called *NEFLIM* that can identify novel fragments of high quality from a pool of decoys with low energy. Using these novel fragments, our method can generate new models with improved quality.

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 tested *NEFLIM* on a benchmark of 30 proteins. The new set of fragments showed better performance when used to predict *de novo* structures. The lowest energy models predicted using our method were closer to native structure than Rosetta for 22 proteins. Following a similar trend the best model among top five lowest energy models predicted using our method were closer to native structure than Rosetta for 20 proteins. In addition, our experiment showed that the CA-RMSD was improved from 5.99 to 5.03 Å on average compared to Rosetta when the lowest energy models were picked as the best predicted models.

4. Conclusion

The use of an estimation of distribution algorithm enabled *EdaFold*$_{AA}$ to reach lower energy levels and to generate a higher percentage of near-native models. *EdaFold*$_{AA}$ 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.

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.

Fiscal Year 2013 List of Publications Resulting from the Use of RICC

[Publication]

1. Simoncini, D., *__Zhang, K. Y. J.__ (2013) Efficient sampling in fragment-based protein structure prediction using an estimation of distribution algorithm. *PLoS ONE*, 8(7): e68954, 1-10. doi:10.1371/journal.pone.0068954

[Proceedings, etc.]

None.

[Oral presentation at an international symposium]

1. The 13th Annual Meeting of the Protein Science Society of Japan, Jun. 12-14, 2013, Tottori, Japan. Invited Speaker, "EdaFold: An Evolutionary Algorithm Based Fragment Assembly Method For *De Novo* Protein Structure Prediction".
2. Center for Quantitative Biology, Peking University, Apr. 23, 2013, Beijing, China. Invited Speaker, "EdaFold: An Evolutionary Algorithm based Fragment Assembly Method for De Novo Protein Structure Prediction".
3. School of Software, Dalian University of Technology, Apr. 26, 2013, Dalian, China. Invited Speaker, "EdaFold: An Evolutionary Algorithm based Fragment Assembly Method for De Novo Protein Structure Prediction".
4. The 12th Meeting of the Asian Crystallographic Association, Dec. 7-10, 2013, Hong Kong. "Improving fragment quality for *de novo* structure prediction".
5. International Conference on Structural Genomics 2013: Structural Life Science, Jul. 29 - Aug. 1, 2013, Sapporo, Hokkaido, Japan. "Error estimation guided rebuilding of de novo models for molecular replacement".
6. 13th Protein Science Society of Japan Annual Meeting, June 12-14, 2013, Tottori, Japan. "EdaFold: A Probabilistic Fragment-based Method for Protein Structure Prediction".
7. 4th International Symposium on Diffraction Structural Biology, May 26-29, 2013, Nagoya, Japan. "Error estimation guided rebuilding of *de novo* models increases the success rate for *ab initio* phasing".

[Others]