RICC Usage Report for Fiscal Year 2014

Project Title:

Protein Folding Prediction Using X-ray Diffraction Data as Constraints

Name:Kam Zhang, Rojan Shrestha, Francois Berenger
Laboratory at RIKEN: Structural Bioinformatics Team, Center for Life Science Technologies

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. Our ability to predict the structures of proteins from their sequences will help us understand how proteins function in the cell.

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. Moreover, new strategies of phasing with low quality models are needed. Our ability to predict protein structure can also benefit the design of new proteins with desired structure and thus new function.

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.

３. Result

We have proposed a method to improve de novo structure prediction by generating new fragment libraries from an initial pool of low energy *de novo* models. It is assumed that fragments in the lowest energy models are most probably the native-like fragments because these fragments are responsible for minimizing the energy. In order to identify these good fragments, fragments from the lowest energy models were clustered. Subsequently, a set of new fragments were selected from the top clusters and then used for a new round of prediction. In a benchmark of 30 proteins, the new set of fragments showed better performance when used to predict *de novo* structures. The lowest energy model predicted using our method was 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 was 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.

We have developed a fragment assembly phasing method that starts from an ensemble of low accuracy *de novo* models, disassembles them into fragments, places them independently in the crystallographic unit cell by molecular replacement, and then reassembles them into a whole structure that can provide sufficient phase information to enable the complete structure determination by automated model building. Tests on ten protein targets have shown that our method can solve structures for eight of these targets, although the predicted *de novo* models cannot be used as templates for successful molecular replacement since the best model for each target is on average more than 4.0 Å away from the native structure. Our method has extended the applicability of the *ab initio* phasing by *de novo* models approach. Our method can be used to solve structures when the best *de novo* models are still of low accuracy.

We have developed a new and rapid computational approach to design proteins with perfect sequence repeat symmetry. As a test case, we have created a 6-fold symmetrical β-propeller protein, and experimentally validated its structure using X-ray crystallography. Each blade consists of 42 residues. Proteins carrying 2 to 10 identical blades were also expressed and purified. Two or three tandem blades assemble to recreate the highly stable 6-fold symmetrical architecture, consistent with the duplication and fusion theory. The other proteins produce different mono-disperse complexes, up to 42 blades (180 kDa) in size, which self-assemble according to simple symmetry rules. Our procedure is suitable for creating nano-building blocks from different protein templates of desired symmetry.

4． Conclusion

The improvement of fragment quality has enabled the de novo modeling method based on fragment assembly to generate better models. The fragmentation and reassembly method has increased the range of de novo models to be used as templates to solve protein crystal structures by molecular replacement. A new method for the design of perfectly sequence symmetric proteins has been proposed and used to design a 6-fold symmetrical β-propeller protein.

5． Schedule and prospect for the future

We will continue to explore new resampling methods for improved de novo structure prediction. The fragmentation and reassembly approach to *ab initio* phasing will be improved to increase its ability to use low quality templates. New computational protein design methods will be proposed to facilitate the design of structures with novel function.

Fiscal Year 2014 List of Publications Resulting from the Use of RICC

[Publication]

1. Shrestha, R., Zhang, K. Y. J. (2015) A fragmentation and reassembly method for *ab initio* phasing. *Acta Cryst.* **D71**, 304-312. doi:10.1107/S1399004714025449.

2. Voet, A. R. D., Noguchi, H., Addy, C., Simoncini, D., Terada, D., Unzai, S., Park, S-Y., Zhang, K. Y. J., Tame, J. R. H. (2014) Computational design of a self-assembling symmetrical β-propeller protein. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 15102-15107. doi: 10.1073/pnas.1412768111.

3. Shrestha, R., Zhang, K. Y. J. (2014) Improving fragment quality for *de novo* structure prediction. *Proteins: Struct., Funct., Bioinf.*, **82**, 2240-2252. DOI: 10.1002/prot.24587.

[Oral presentation at an international symposium]

1. The Eisenberg Symposium, Feb. 22-23, 2014, The University of California at Los Angeles, Westwood, Los Angeles, USA. Invited Speaker, "An Evolutionary Algorithm Based Fragment Assembly Method For De Novo Protein Structure Prediction".
2. 11th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction, Dec. 7-10, 2014, Riviera Maya, Mexico, Poster presentation. David Simoncini, Arnout R.D. Voet, Kam Y. J. Zhang, "RosEda: Combining Rosetta AbInitio with an Estimation of Distribution Algorithm".
3. 11th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction, Dec. 7-10, 2014, Riviera Maya, Mexico, Poster presentation. Rojan Shrestha, Kam Y. J. Zhang, "NEFILIM: improving fragment quality for protein structure prediction".
4. The 23rd Congress and General Assembly of the International Union of Crystallography, Aug. 5-12, 2014, Montreal, Canada, Poster presentation. Rojan Shrestha, Kam Y. J. Zhang, "Error estimation guided rebuilding of *de novo* models for *ab initio* phasing".
5. The 2014 International Biophysics Congress, Aug. 3-7, 2014, Brisbane, Australia, Poster presentation. Arnout RD Voet, Hiroki Noguchi, Christine Addy, Daiki Terada, David Simoncini, Kam Y. J. Zhang, Jeremy RH Tame, "Computational design of a self-assembling symmetrical β-propeller".