Project Title:

Use resampling techniques to improve de novo structure prediction based on fragment assembly

Name:Kam Zhang, Rojan Shrestha, David Simoncini
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 prediction of 3D structures from amino acid sequences directly is a computationally challenging task. One of the most effective approaches uses fragment-assembly. We have developed two methods that use resampling techniques to improve the fragment-assembly based de novo structure prediction. Our first method uses an evolutionary algorithm to focus the sampling on those fragments that are most productive and has increased the efficiency of the sampling and generated better quality structures. Our second method aims at identifying new and better quality fragments from the initial pool of generated models and uses them for the subsequent round of model generation. We propose to test our methods in the 11th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP11). The results from this prospective experiment will provide invaluable information as how to further improve our methods. Both of our methods require a large amount of computational resources and this can only be achieved on RICC.

2. Specific usage status of the system and calculation method

Our first method is called RosEda, which is an iterative algorithm using the concept of Estimation of Distribution Algorithm to gather information between initially independent predictions. At the first iteration, there is a uniform distribution over the library and every fragment has the same probability of being selected. At each subsequent iterations, a fraction of the lowest energy models is selected as a sample set. To compute the energy, all the models are relaxed in their all-atom representation via Rosetta Relax protocol. The probabilities of selecting fragments for insertion during subsequent iterations are modified according to the observed distribution of fragments used in the sample set.

Our second method is called NEFILIM, which take advantage of low quality models predicted in the initial round to identify better fragments. The NEFILIM protocol used in CASP11 has three major steps – initial model generation, fragment generation, and final model generation. First, the initial-run starts with a target sequence and fragments generated using Rosetta. Rosetta was used to generate 120,000 all-atom models in the initial-run. In fragment generation step, one thousand low energy models were selected and these models were fragmented into three-residue and nine-residue respectively. The fragments were clustered for both types of fragments and then twenty-five fragments were selected from the top five clusters. Thirty thousand all-atom models were generated in the new-run. We selected one thousand low energy models from two runs, which was 150,000 models in total. Average pairwise residue distant score (APRDS) among 1000 low energy models were computed for model selection. Twenty low energy models were taken for visual inspection and five models were selected for final submission.

3. Result

We have generated between 90,000 and 240,000 models using RosEda depending on the sequence length and available resources. Models were submitted for 54 targets out of the 55 human prediction targets. The method was designed for de novo modeling when no homologous structures are available. It belongs to the Free-Modeling (FM) category. When judged by the assessors' formula, the best rank that RosEda achieved is 3 out of 139 groups in the FM category.

For structure prediction with NEFILIM, we have submitted predictions for 44 targets out of the 55 human prediction targets. The best rank that NEFILIM achieved is 4 of 139 groups in the FM category when measured according to the assessors' formula.

4. Conclusion

We have achieved very encouraging results with our fragment-based de novo modeling methods, RosEda and NEFILIM, in the CASP11. Our methods were able to predict models with high accuracy even for some targets that belong to the TM category, although no homologous structure templates were used explicitly. Our participation in this CASP11 exercise also revealed that the importance of being able to parse a target into FM or TM category based on its amino acid sequence. It has also shown that the ability to handle multi-domain or oligomeric structures is critical for de novo modeling methods.

5. Schedule and prospect for the future

Based on the experience obtained from our participation in CASP11, we will further improve the structure prediction accuracy of our de novo modeling methods. We will focus more on developing methods to handle multi-domain and oligomeric structures in order to expand the utility of our structure prediction methods.

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

[Publication]

1. 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. 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".
2. 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".