Project Title:

Protein Folding Prediction Using X-ray Diffraction Data as Constraints

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

X-ray crystallography is the principle method of determining 3D structures of proteins. It uses the diffraction phenomenon caused by the interaction of crystals with X-rays. Only the amplitudes of the diffraction can be measured experimentally but not their phases. However, both the amplitudes and phases are required to reconstruct the crystallographic image in order to reveal the atomic positions in the crystal. Phase retrieval is therefore of crucial importance in macromolecular structure determination.

The *ab initio* phasing is one of remaining challenges in protein crystallography. Although the molecular replacement method can be used to solve the phase problem when a homologous model is available, labor intensive and costly experimental phasing methods have to be carried out for proteins with novel folds. It has been demonstrated recently that computationally predicted *de novo* models have reached high enough accuracy to solve the phase problem *ab initio*.

There are several bottlenecks in the "*ab initio* phasing with *de novo* models" method. First is the inaccuracies in the free energy functions used, and therefore a purely free energy based criteria to identify best predicted structure is insufficient. The second is conformation space sampled is not enough, and a more efficient search algorithm is needed. The third is that the use of molecular replacement during the folding simulation increases the computational demand dramatically, and this requires new and fast methods of evaluating the fit of a decoy to the given diffraction data.

2. Specific usage status of the system and calculation method

To address the conformation sampling challenge, we have developed a method that uses the Estimation of Distribution Algorithm to increase the conformational sampling efficiency. This method is implemented in a program called EdaFold for protein structure prediction which uses Rosetta 3.2 as a library. The parallel version of this program is running on RICC using MPI. The program also has a dependency on the Evolving Object C++ library for the representation of data and for statistics.

The EdaFold program is running multiple predictions in parallel, allowing communications between the different protein models generated in order to improve the quality of subsequent models. The communication between processes allows us to estimate the optimal distribution over the fragments of proteins used to build the models.

We are currently working on a surface descriptor useful to analyze the surface of proteins. Proteins have an extremely strange surface landscape. For example, there is no such thing as a sea-level on a protein so some "classic" surface descriptors cannot be employed for the surface of proteins. The descriptor we propose to use is computed in a parameter-less fashion. This descriptor has several potential applications in bioinformatics.

We have also developed a new approach for improvement of overall conformation quality by removing noise introduced due to the local residues. After removing the noise in local residues, the improved models were optimized using Rosetta 3.2. The resultant models have reached the level of accuracy necessary to solving crystallographic phase problem.

We are also working on developing methods for model quality assessment (MQA) in protein structure prediction. We used average pairwise (AP) method in APOLLO (which is designed to calculate 4 types of pairwise score in group of predicted models), and single method using OPUS and Model Evaluator (which are designed to calculate MQA score for each predicted model). To find whether hybrid method shows better result, Z-score of AP method result was applied to the result of single method under various combination rate.

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 (successful) 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 further accelerated our protein decoys exact clustering software by integrating the Quaternion-based Characteristic Polynomial (QCP) method to compute RMS distances faster. The tool is used by protein folding researchers and possibly Molecular Dynamics (MD) practitioners in order to analyze trajectories produced by MD simulations. This new software is released as open source and available from our laboratory's homepage (http://www.riken.jp/zhangiru/en_software.html) under the "Durandal with QCP" section.

Our proposed 3D structure local refinement method has reduced local noise in computationally generated coarse-grained models. Therefore subsequent models after minimization using Rosetta all-atom energy were used for estimation of the phases. Our current method is tested on ten structure factor sets which were unsuccessful using our previous approach. Out of ten, seven structure factors were successfully phased with improved *de novo* models.

We tested all the models which were submitted to CASP9 from various protein structure prediction servers. The MQA result using AP method showed correlation average 0.89 (Pearson's r) with GDT scores with crystal structure. Single method of OPUS and Model evaluator showed 0.6 on average. The result of hybrid method with optimal rate between AP method and single methods showed better correlation than just using AP method. The gap between best result of AP method and that of hybrid method was +0.162. The result of hybrid method is always better in any rate of applying when the median of single score is higher than that of AP score.

4. Conclusion

We have found that allowing communications

between protein models during prediction runs can increase the percentage and the quality of near-native protein models. Our EdaFold method, which relies on the estimation of the distribution over fragments of proteins used to build models can be embedded in any fragment-based approach and enhance its performance.

We have found that incorporating the Quaternion-based Characteristic Polynomial method into our protein decoys clustering software has increased its speed between 13% and 27% compared to the previous version.

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.

The result of AP method for MQA is usually better than that of single MQA method. We found, if the median of AP score is lower than that of single score, the MQA result of hybrid method is better than any single method, in any applying rate between these two.

5. Schedule and prospect for the future

We plan to extend our EdaFold method from using the coarse-grained energy function for the estimation of distribution to using the all-atom energy function. We believe that this will provide improved estimation and increase the sampling efficiency.

We also plan to use the principle of Estimation of Distribution Algorithm to improve the sampling efficiency of side-chains. Our previous work has been improving the sampling of the main-chains. The sampling of side-chain rotamers presents a different challenge. However, we believe that our method should also work for side-chain sampling.

We plan to use the surface descriptor that we have developed to detect pockets on protein surfaces that are suitable for small molecule ligand binding. There are several methods for protein pocket detection. All of them use several parameters. This makes them less attractive to users. We will develop a parameter-free method for the detection of protein pockets.

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 you wish to extend your account, provide usage situation (how far you have achieved, what calculation you have completed and what is yet to be done) and what you will do specifically in the next usage term.

We wish to extend our account to the next fiscal year. We have made significant achievement during this fiscal year. Specifically, we have developed a new and efficient conformation sampling method called EdaFold based on the Estimation of Distribution Algorithm. We have demonstrated that EdaFold could obtain better quality decoys and with higher probability than Rosetta. We have developed a method that could identify regions in the predicted structure where the accuracy is low and should be improved. These low accuracy regions identified are subject to more intense re-sampling to improve their predicted structures. We found that this protocol could improve the RMSD to native by about 0.5Å on average for all the structure tested. This improved structure has led to improved success rate in solving the X-ray crystallographic phase problem by

molecular replacement using these improved *de novo* models.

There is much to be done towards achieving our overall goal of solving the crystallographic phase problem using de novo models. The specifics that we plan to achieve in the next usage term are briefly sketched in the previous section.

7. If you have a "General User" account and could not complete your allocated computation time, specify the reason.

We have not completely used our allocated computation time in this fiscal year. One of the reasons is that the RICC was not available immediately after the earthquake and was available sporadically for a while due to the electricity blackout.

8. If no research achievement was made, specify the reason.

N/A.

Fiscal Year 2011 List of Publications Resulting from the Use of RICC

[Publication]

1. Berenger, F., Shrestha, R., Zhou, Y., Simoncini, D., Zhang, K. Y. J. (2012) Durandal: fast exact clustering of protein decoys. *J. Compu. Chem.*, **33**, 471-474.
2. Shrestha, R., Berenger F., Zhang, K. Y. J. (2011) Accelerating *ab initio* phasing with *de novo* models. *Acta Cryst.* **D67**, 804-812.
3. Berenger F., Zhou Y., Shrestha, R., Zhang, K. Y. J. (2011) Entropy-accelerated exact clustering of protein decoys. *Bioinformatics*, **27**, 939-945.

[Proceedings, etc.]

If a publication does not contain the acknowledgement, please provide the reason for the missing of acknowledgement

[Oral presentation at an international symposium]

[Others]