RICC Usage Report for Fiscal Year 2010

Project Title:

Protein Folding Prediction Using X-ray Diffraction Data as Constraints

Name :○Kam Zhang, Rojan Shrestha, Francois Berenger
Ashutosh Kumar, Kamlesh Kumar Sahu

Affiliation : Zhang Initiative Research Unit Advanced Science Institute,Wako Institute

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

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*. This “*ab initio* phasing with *de novo* models” method first generates a huge number of *de novo* models and then selects some lowest energy models to solve the phase problem using molecular replacement. The number of models generated was about 10^5 and it took over 100 CPU days to complete even for small size proteins with about 130 amino acids or less. This amount CPU time has limited the utility of this method.

The current “*ab initio* phasing with *de novo* models” has been carried out on selected models after the generating huge number of models. We seek to develop a new method to carry out the phasing during the folding process. Our approach could significantly reduce the computing time required to perform the “*ab initio* phasing with *de novo* models”. Instead of performing molecular replacement after the completion of all models, we initiate molecular replacement during the course of each simulation. The molecular replacement solutions are closely monitored and all subsequent simulations are terminated once some solutions above a threshold have been obtained.

We would also like to develop method that parallelize the job execution process and maximize the efficient utilization of computing resources so that we could tackle the vast conformational search problem facing us.

Another challenge is the inaccuracies in the free energy function used during protein folding. Consequently, the lowest free energy predicted structure may not be the best structure (i.e. the closest to the native structure). We would like to develop more sensitive method to identify the best predicted structure among many decoys.

２. Specific usage status of the system and calculation method

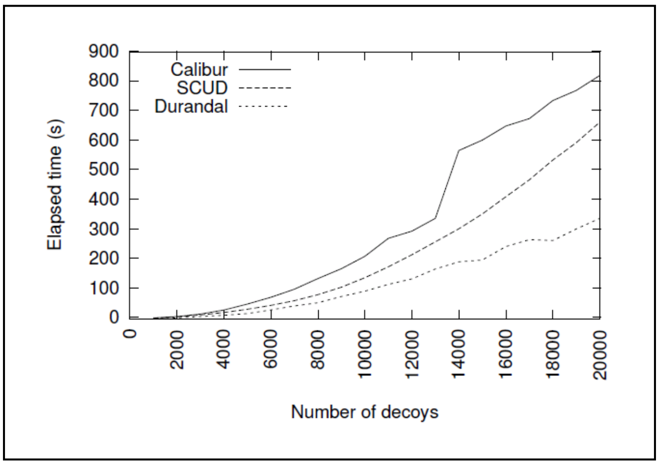
We have used Rosetta 3.0 (which is designed to predict the 3D structure of a protein given its amino acid sequence using fragment assembly approach) to perform the protein folding simulation and Phaser 2.1.4 (which is a program for phasing macromolecular crystal structures by molecular replacement using the maximum likelihood) for molecular replacement calculations. In order to initiate molecular replacement during the course of each protein folding simulation, we have converted a standalone Phaser 2.1.4, into an object-oriented version that is callable via a library and incorporated into Rosetta 3.0 to produce a modified program named as RosettaX.

Computational method for protein folding requires huge computational power and CPU-time due to the astronomical number of conformations that have to be searched and evaluated. Typically, the numbers of conformations have to be sampled is in the order of hundred thousand to millions for each protein. When initiating molecular replacement during the course of each protein folding simulation, the computing time required is further increased dramatically. In order to make computation more efficient, we initiate Phaser only during the six distinctive stages of all-atom refinement protocol in Rosetta. We have also developed a statistically derived and experimentally validated upper and lower thresholds for Phaser score to predict the likelihood of success or failure of each decoy in molecular replacement. We then closely monitor each spawned folding simulation process in a parallelized execution environment. We terminate a simulation when the Phaser score of a decoy it generated is below the lower threshold and above the upper threshold. When sufficient number of decoys scored higher than the upper threshold have obtained, the entire simulation run is terminated.

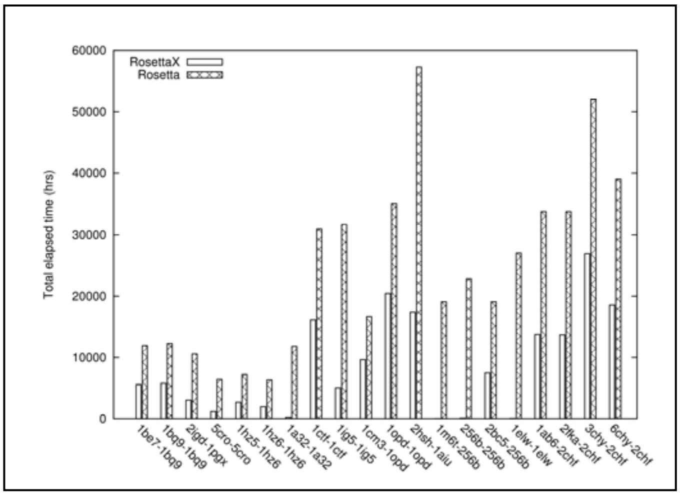
３. Result

The Rosetta approach to protein structure prediction using fragment assembly is a highly parallel process. Many independent folding simulations are running simultaneously on multiple CPUs. These simulations can be considered as a "Bag-of-Tasks (BoT)". This BoT can be handled by a job scheduler which requires a pre-specified number of CPUs as resources. Instead of waiting for all the pre-specified number of CPUs available to start the BoT job execution and in order to efficiently utilize computing resources, we have developed PAR, a scalable, dynamic, parallel and distributed execution engine. PAR works in a *pull* mode and inspired by desktop grid platforms. Workers *join* the computation and can be added dynamically at run-time; the server delivers tasks to workers available at a given moment. Our RosettaX runs can be executed through PAR without pre-specify the number of available CPUs. The job execution dynamically allocates the available CPUs at runtime. This has significantly increased the efficiency of our utilization of computing resource on RICC.

In order to identify the best predicted structure among many decoys more accurately and to avoid the pitfalls of using free energy only, we have developed a new method of using clustering to identify the best predicted structures. We proposed a method using propagation of geometric constraints to accelerate exact clustering, without compromising the distance measure. Our method can be used with any metric distance. Metrics that are expensive to compute and have known cheap lower and upper bounds will benefit most from the method. We compared our method's accuracy against published results from the SPICKER clustering software on 40 large decoy sets from the I-TASSER protein folding engine. We also performed some additional speed comparisons on six targets from the 'semfold' decoy set. In our tests, our method chose a better decoy than the energy criterion in 25 out of 40 cases versus 20 for SPICKER. Our method also was shown to be consistently faster than another fast software performing exact clustering named Calibur. In some cases, our approach can even outperform the speed of an approximate method. The speed comparison between Durandal, Calibur and SPICKER is shown in the figure below.



We have tested RosettaX on a benchmark dataset of 20 proteins. We found that our method is 141x faster than the conventional approach (Rosetta). The comparison between RosettaX and Rosetta is shown in the figure below.



4. Conclusion

We have found that in most cases molecular replacement solutions were determined soon after the coarse-grained models were turned into full atom representations. We have also found that all-atom refinement could hardly change the models sufficiently to enable successful molecular replacement if the coarse-grained models were not very close to the native structure. Therefore, the conformation sampling of coarse-grained models still remains a major challenge in predicting structures accurate for molecular replacement.

Ab initio phasing with *de novo* models during protein folding is efficient way for phasing in the absence of homologous model. Phasing while folding parsimoniousely uses computation resources and subsequently reduces computation time dramatically.

5. Schedule and prospect for the future

This is described in our new RIKEN Supercomputer System Grant Request Form.

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.

Yes, we wish to extend our account, but we have to write a new application. The details are described in our new RIKEN Supercomputer System Grant Request Form.

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

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

[Publication]

1. Berenger F., Zhou Y., Shrestha R., Zhang K. Y. J. (2011) Entropy-accelerated exact clustering of protein decoys. *Bioinformatics*, doi: 10.1093/bioinformatics/btr072
2. Berenger F., Coti C. and Zhang K. Y. J. (2010) PAR: A PARallel And Distributed Job Crusher. *Bioinformatics*, **26**, 2918–2919.