Usage Report for Fiscal Year 2016

Project Title:

Protein structure prediction and design

Name:

Kam Zhang
Pierre Thevenet
Daiki Terada
Aditya Padhi

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

Proteins are fundamental components of life and among the most studied macro-molecules, mainly because their malfunctions often lead to diseases. Being able to predict the three-dimensional structure of proteins is critical in order to better understand the relationships between sequence, structure and function. Still, while tens of millions of protein sequences have been discovered, protein structure prediction remains a challenging problem and the Protein DataBank gathers no more than one hundred thousands structures.

Conformational search space exploration remains a major bottleneck for protein structure pre-diction methods. Population-based meta-heuristics typically enable the possibility to control the search dynamics and to tune the balance between local energy minimization and search space exploration. EdaFold is a fragment-based approach that can guide search by periodically updating the probability distribution over the fragment libraries used during model assembly. We propose to implement the EdaFold algorithm as a Rosetta protocol and make our algorithm more readily accessible to the scientific community.

2. Specific usage status of the system and calculation method

The Rosetta AbRelax protocol made the fragment approach for protein structure prediction popular due to its success in the CASP experiments. The method is based on the assembly of short structural protein fragments retrieved from the PDB. The fragments are assembled during a coarse grained phase where protein side chains are represented by their centroids. The protein models are then refined in a fine grained all atom phase. The fragment assembly phase is divided into several stages, each of which adds terms to the energy function or modifies the terms' weights.

Our EdaRose method implements two different strategies to select the models used for the estimation of distribution. In the first strategy, the models are selected according to their energy. An energy threshold is defined in order to control the fraction of protein models upon which the estimation of distribution is made. The second strategy uses two criteria: energy evaluation and structural dissimilarity with other selected models.

3. Result

We tested the performance of EdaRose on a benchmark of 20 proteins of various length and structural classes. Fragment libraries were

generated (excluding homologues) using the Robetta Server. 30, 000 predictions were made with EdaRose, divided in 6 iterations producing 5, 000 models each. EdaRose outperforms Rosetta AbRelax on a majority of the targets in this benchmark. It obtains the highest number of first, best and average top 10 predictions. EdaRose also gets the lowest average values over the whole benchmark according to these three criteria.

4. Conclusion

Population-based meta-heuristics are powerful optimization methods typically employed on problems with huge search space and rugged fitness landscape. Whereas the protein structure prediction problem possesses these characteristics, single solution methods are still the norm. The success of our approach, and generally of all population-based meta-heuristics, relies on its ability to efficiently balance exploration of the search space and exploitation of the good solutions that were discovered so far. Therefore, meta-heuristics usually include some mechanisms that allow to tune the exploration / exploitation trade-off. In our case, the selection of candidate models that are used during the estimation of distribution represents a nice spot to inflect the search dynamics.

5. Schedule and prospect for the future

EdaRose is implemented as Rosetta protocols, and share a fraction of source code related to communication between runs and distribution updates. This fraction of code provides the opportunity to develop many more Rosetta protocols by implementing new update strategies. In the future, we plan to update strategies relying on protein features, or using model quality assessment methods. In the ideal case, the update strategy would allow to find one model per local minimum attraction basin and get as deep as possible in each basin.

6. If no job was executed, specify the reason.

N/A.

Usage Report for Fiscal Year 2016

Fiscal Year 2016 List of Publications Resulting from the Use of the supercomputer

[Publication]

1. Simoncini, D., Schiex, T. and Zhang, K. Y. J. (2017) Balancing exploration and exploitation in population-based sampling improves fragment-based de novo protein structure prediction. *Proteins: Struct., Funct., Bioinf.*, DOI:10.1002/prot.25244.

[Oral presentation at an international symposium]

1. Computer Aided Drug Design Workshop, 2016 (CADD 2016), Oct. 30 – Nov. 2, 2016, Langkawi, Malaysia. Invited Speaker, “Protein Design”.
2. 24th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB2016), July 8-12, 2016, Orlando, Florida, USA. Invited Speaker, “RE_3Volutionary Design of Symmetric Proteins That Biomineralize Nanocrystals”.