Usage Report for Fiscal Year 2021

Project Title: Protein Structure Prediction and Design

Name: ○Kam Zhang (1), Aditya Padhi (1), Rahul Kaushik (1), Chun Lai Tam (1), Francois Berenger (1)

Laboratory at RIKEN:
(1) Laboratory for Structural Bioinformatics, Center for Biosystems Dynamics Research

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

An accurate estimation of the quality of protein model structures typifies as a cornerstone in protein structure prediction regimes. Despite the recent groundbreaking success in the field of protein structure prediction, there are certain prospects for the improvement in model quality estimation at multiple stages of protein structure prediction and thus, to further push the prediction accuracy.

2. Specific usage status of the system and calculation method

In order to improve model quality estimation, a novel approach, named ProFitFun, for assessing the quality of protein models is proposed by harnessing the sequence and structural features of experimental protein structures in terms of the preferences of backbone dihedral angles and relative surface accessibility of their amino acid residues at the tripeptide level. The proposed approach leverages upon the backbone dihedral angle and surface accessibility preferences of the residues by accounting for its N-terminal and C-terminal neighbors in the protein structure. These preferences are employed to evaluate protein structures through a machine learning approach and tested on an extensive dataset of diverse proteins.

3. Result

The approach was extensively validated on a large test dataset (n = 25,005) of protein structures, comprising 23,661 models of 82 non-homologous proteins and 1,344 non-homologous experimental structures. Additionally, an external dataset of 40,000 models of 200 non-homologous proteins was also used for the validation of the proposed method. Both datasets were further employed for benchmarking the proposed method with four different state-of-the-art methods for protein structure quality assessment. In the benchmarking, the proposed method outperformed some state of the art methods in terms of Spearman's and Pearson's correlation coefficients, average GDT-TS loss, sum of z-scores, and average absolute difference of predictions over corresponding observed values. The high accuracy of the proposed approach promises a potential use of the sequence and structural features in computational protein design.

4. Conclusion

A reliable quantification of the quality of protein models is pivotal to the success of protein structure prediction regimes. We have harnessed the sequence and structural features of backbone dihedral angles and relative surface accessibility preferences of amino acids by accounting for their neighboring residues in the protein structures and developed a novel method for quantifying the accuracy of protein tertiary structures. The developed method, ProFitFun, has outperformed some of the state-of-the-art methods when benchmarked on a comprehensive dataset of about 65,000 proteins structures of non-homologous proteins. The easy implementation with minimal additional requirements, time and computing efficiencies, and portability to Linux and Windows OS of ProFitFun impart some additional advantages over other methods. Likewise, the high accuracy of this sequence and structural features based method

promises a potential application of these features in computational protein design and other sub-areas of computational protein folding.

5. Schedule and prospect for the future

We plan to further improve the methodology to increase its accuracy in model quality assessment. First, we will develop an integrated consensus approach based protein structure evaluation method for determining the quality of decoys to signalize the best or near-best model structure. Second, we will develop a residue-wise protein structure evaluation method for identifying incorrectly modeled regions in predicted modeled structures. Finally, we will use these new model quality assessment methods in the design of novel proteins.

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

Fiscal Year 2021 List of Publications Resulting from the Use of the supercomputer

[Paper accepted by a journal]

1. Tam, C., Zhang, K. Y. J. (2022) FPredX: interpretable models for the prediction of spectral maxima, brightness and oligomeric states of fluorescent proteins. *Proteins: Struct., Funct., Bioinf.*, **90**, 732-746. https://doi.org/10.1002/prot.26270.
2. Kaushik, R., Zhang, K. Y. J. (2022) ProFitFun: A Protein Tertiary Structure Fitness Function for Quantifying the Accuracies of Model Structures. *Bioinformatics*, **38**, 369–376. https://doi.org/10.1093/bioinformatics/btab666.
3. Tam, C., Kumar, A., Zhang, K. Y. J. (2021) NbX: Machine Learning Guided Re-ranking of Nanobody-Antigen Binding Poses. *Pharmaceuticals*, **14**, 968. https://doi.org/10.3390/ph14100968.
4. *Kumar, N., Kaushik, R., Tennakoon, C., Uversky, V. N., Mishra, A., Sood, R., Srivastava, P., Tripathi, M., Zhang, K. Y. J., Bhatia, S. (2021) Evolutionary signatures governing the codon usage bias in coronaviruses and their implications for viruses infecting various bat species. *Viruses*, **13**, 1847. https://doi.org/10.3390/v13091847.
5. Yagi, S., Padhi, A. K., Vucinic, J., Barbe, S., Schiex, T., Nakagawa, R., Simoncini, D., Zhang, K. Y. J., Tagami, S. (2021) Seven amino acid types suffice to create the core fold of RNA polymerase. *J. Am. Chem. Soc.*, **143**, 15998–16006. https://doi.org/10.1021/jacs.1c05367.
6. Padhi, A. K., Kumar, A., Haruna, K., Sato, H., Tamura, H., Nagatoishi, S., Tsumoto, K., Yamaguchi, A., Iraha, F., Takahashi, M., Sakamoto, K., Zhang, K. Y. J. (2021) An integrated computational pipeline for designing high-affinity nanobodies with expanded genetic codes. *Brief. Bioinformatics*, **22**, 1-17. https://doi.org/10.1093/bib/bbab338.
7. Kukimoto-Niino, M., Katsura, K., Kaushik, R., Ehara, H., Yokoyama, T., Uchikubo-Kamo, T., Nakagawa, R., Mishima-Tsumagari, C., Yonemochi, M., Ikeda, M., Hanada, K., Zhang, K. Y. J., Shirouzu, M. (2021) Cryo-EM structure of the human ELMO1-DOCK5-Rac1 complex. *Sci. Adv.*, **7**, eabg3147. https://doi.org/10.1126/sciadv.abg3147.
8. Bhattacharya, S., Nautiyal, A. K., Bhattacharya, R., Padhi, A. K., Junghare, V., Bhambri, M., Dasgupta, D., Zhang, K. Y. J., Ghosh, D., Hazra, S. (2021) A comprehensive characterization of novel CYP-BM3 homolog (CYP-BA) from *Bacillus aryabhattai*. *Enzyme Microb. Technol.*, **148**, 109806. https://doi.org/10.1016/j.enzmictec.2021.109806.
9. Kumar, N., Kaushik, R., Tennakoon, C., Uversky, V. N., Longhi, S., Zhang, K. Y. J., Bhatia, S. (2021) Insights into the evolutionary forces that shape the codon usage in the viral genome segments encoding intrinsically disordered protein regions. *Brief. Bioinformatics*, **22**, 1-12. https://doi.org/10.1093/bib/bbab145.

[Conference Proceedings]

[Oral presentation]

[Poster presentation]

[Others (Book, Press release, etc.)]