Usage Report for Fiscal Year 2022

Project Title: Protein Structure Prediction and Design

Name: ○Kam Zhang (1), Aditya Padhi (1), Rahul Kaushik (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

The structural information of a protein is pivotal to comprehend its functions, protein-protein and protein-ligand interactions. There is a widening gap between the number of known protein sequences and that of experimentally determined structures. The protein structure prediction has emerged as an efficient alternative to deliver the reliable structural information of proteins. However, it remains a challenge to identify the best model among the many predicted by one or a few structure prediction methods.

2. Specific usage status of the system and calculation method

In order to improve model quality estimation, a neural network model, ProFitFun-Meta, was developed by implementing a replicable training scheme on the Training Dataset (n = 69,243) with a maximum number of 500 iterations through a multi-layer perceptron algorithm with a backpropagation approach having 5 fully connected hidden layers of 100 neurons. The combinations of different activator and solver functions were used to identify the best parameters. The cross validation (10-fold) using different activation and solver functions is assessed. The best combination of activation function for the hidden layers and solver function for the weight optimization was selected as a function of Mean Square Error (MSE), Mean Absolute Error (MAE) and Pearson's Correlation Coefficient (r) for the GDT-TS score prediction in the cross validation. Notably, the combination of Logistic activator (the logistic sigmoid function) and Adam's solver (stochastic gradient-based optimizer) function showed the best evaluation statistics in cross validation for the development of ProFitFun-Meta.

3. Result

The evaluation of the ProFitFun-Meta for the prediction of GDT-TS was performed on the basis of the statistics on the Test Dataset (n = 26,604) and an External Dataset (40,000 model structures of 200 non-homolog proteins). The ProFitFun-Meta demonstrated an improved performance in terms of Pearson's correlation coefficient ($r = 0.75$), Spearman's correlation coefficient ($\rho = 0.70$), absolute loss ($d = 0.128$), and GDT-TS loss ($g = 0.132$) over all other methods used in the benchmarking. The ProFitFun-Meta performed significantly better over its antecedent ProFitFun ($r = 0.70$, $\rho = 0.65$, $d = 0.139$, and $g = 0.144$), that indicated the complementary effect of additional features in ProFitFun-Meta in achieving better accuracy for predicted GDT-TS scores of protein model structures. The significance of ProFitFun-Meta performance was statistically analyzed by a two-sample Kolmogorov-Smirnov test. The statistical analysis supported the significant difference (p-value ≤ 0.001) in performance of ProFitFun-Meta over other methods used in the benchmarking

4. Conclusion

We have integrated the structural information of experimentally characterized proteins, in terms of backbone dihedral angles and surface accessibility preferences of residues, spatial properties of protein structures for quality estimation (side chain and backbone clashes, rotamer orientations, and Cα geometry), and potential of model structures to mimic the

experimentally solved protein structures by accounting for the deviations of covalent interactions, dihedral angles and overall interactions. These parameters were unified together through a neural network based machine learning model, ProFitFun-Meta, to deliver a highly precise quality evaluation of the protein model structures. The accuracy of ProFitFun-Meta is validated on an extensive dataset of ~65,000 model structures of non-homologous proteins.

A better knowledge of structural and functional features of proteins is pivotal to various therapeutic developments in identifying novel drug targets. ProFitFun-Meta can furnish a reliable quantification of the structural quality of the predicted model structures that are generated through different protein structure prediction methodologies. For instance, the reliable scoring of model structures can expedite an optimal utilization of structural models for computer aided drug discovery regimes, ligand binding studies, structure to function characterization, and various protein-protein interaction and mutagenesis studies. The accuracy of ProFitFun-Meta can be easily harnessed by incorporating it into the computational pipelines for protein modeling and protein design.

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.

Usage Report for Fiscal Year 2022

Fiscal Year 2022 List of Publications Resulting from the Use of the supercomputer

[Paper accepted by a journal]

1. Kaushik, R., Zhang, K. Y. J. (2022) An Integrated Protein Structure Fitness Scoring Approach for Identifying Native-Like Model Structures. *Comput. Struct. Biotechnol. J.*, **20**, 6467–6472. https://doi.org/10.1016/j.csbj.2022.11.032.
2. 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.
3. 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.