Usage Report for Fiscal Year 2019

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

The infinitesimally small sequence space naturally scouted in the millions of years of evolution suggests that the natural proteins are constrained by some functional prerequisites and should differ from randomly generated sequences. A primary and a very crucial step of novel protein design involves the computational identification or generation of potential protein sequences having a considerably high probability of mimicking the naturally occurring proteins and eventually folding into a compact structure.

We have developed a sequence and secondary structure-based fitness scoring function to identify potentially foldable/designable protein sequences by differentiating them from non-natural and random protein sequences. The presented fitness function implements the competency scores derived from sequences and corresponding secondary structures of well-characterized known protein domains and amino acid composition constrained computationally generated non-natural protein sequences. The scoring function classifies a query protein sequence into foldable (natural protein) or non-foldable (non-natural and/or random protein) depending on its competency scores compared with natural and non-natural protein sequences.

2. Specific usage status of the system and calculation method

For all the protein sequences in the natural proteins reference dataset, tripeptides frequencies are calculated for all possible 8000 combinations. Also, individual amino acid residues occurrence frequencies are calculated from natural proteins reference dataset. It may be noted that the natural protein reference dataset represents all the possible combinations at tripeptides level sufficiently, encompassing more than 8 million tripeptides

The conditional probabilities of all tripeptides as calculated are further used to compute a percentage sequence competency score (CS-Score) at individual residue level by normalizing the conditional probabilities with the maximum conditional probability in all combinations of tripeptides.

The sequence and sequence and secondary structure scoring libraries (CS-scores and CSS-Scores) are used to calculate average competency scores for individual sequences in natural proteins reference dataset of 41132 proteins. For all the computationally generated protein sequences in non-natural protein reference dataset, the average competency scores for individual sequences are calculated by using tripeptide-based CS-Scores and CSS-Scores.

3. Result

The performance of CS- and CSS-Scores is evaluated on a test dataset of natural and non-natural proteins (17,626 proteins each, as mentioned in section 2.3). Additionally, a dataset of ~57,000 unique natural proteins (clustered at 40% sequence identity) of sequence length varying from 50 to 700 residues from UniProtKB is selected after excluding all the natural proteins of SCOPe database (58758 proteins). Further, other datasets of ~57,000 computationally generated proteins each with and without amino acid composition constraint (all amino acid with equal occurrence probability) are considered for

quantifying the ability of competency scores in differentiating natural proteins from non-natural proteins.

The proposed fitness function is extensively validated on a dataset of about 210,000 natural and non-natural protein sequences and benchmarked with existing methods for differentiating natural and non-natural proteins. The high sensitivity, specificity and percentage accuracy (0.81, 0.95 and 91% respectively) of the fitness function demonstrates its potential application for sampling the protein sequences with higher probability of mimicking natural proteins. Moreover, the four major classes of proteins (α proteins, ß proteins, α/ß proteins and α+ß proteins) are separately analyzed and ß proteins are found to score slightly lower as compared to other classes. Further, an analysis of about 250 designed proteins (adopted from previously reported cases) helped define the boundaries for sampling the ideal protein sequences. The protein sequence characterization aided by the proposed fitness function could facilitate the exploration of new perspectives in the design of novel functional proteins.

4. Conclusion

The proposed scoring function is extensively validated on a dataset of about 210,000 natural and non-natural protein sequences and benchmarked with existing methods for differentiating natural and non-natural proteins. The high sensitivity, specificity and percentage accuracy (0.81, 0.95 and 91% respectively) of the scoring function demonstrates its potential application for sampling the protein sequences with higher probability of mimicking natural proteins. Also, the four major classes of proteins (α proteins, ß proteins, α/ß proteins and α+ß proteins) are separately analyzed and ß proteins are observed to scoring slightly on the lower side as compared to other classes. Further, an analysis of about 250 designed proteins (adopted from previously reported cases) helped in defining the boundaries for sampling the ideal protein sequences which may prove advantageous in computational protein design regimes.

5. Schedule and prospect for the future

Our scoring functions that implement sequence and secondary structural information at tripeptide level have demonstrated good performance in differentiating natural from non-natural protein and are found to be superior to several published methods. However, there are still room for improvement. We plan to investigate the incorporation of the dihedral angle information of the amide bond to improve the accuracy of the scoring functions. We also plan to use Deep Learning methodologies to obtain a method with even better capability for the identification of native proteins based only on sequences.

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

Usage Report for Fiscal Year 2019

Fiscal Year 2019 List of Publications Resulting from the Use of the supercomputer

[Paper accepted by a journal]

1. Simoncini, D., Zhang, K. Y. J., Schiex, T., Barbe, S. (2019) A Structural Homology Approach for Computational Protein Design with Flexible Backbone. *Bioinformatics*, **35**, 2418-2426.

[Conference Proceedings]

[Oral presentation]

1. RIKEN BDR Symposium, Mar. 2-4, 2020, Kobe, Japan. Invited speaker, “Evolution-inspired computational design of symmetric proteins”.

[Poster presentation]

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