Balanced prediction of protein secondary structure

Proteins by its three-dimensional (3D) structure play significant roles in the biological processes within the cell. Thus, it is important to understand the 3D structures of a protein. However, accurate prediction of the 3D structure of a protein relies on precise secondary structure (SS) prediction. The SS defines the local spatial organization of protein's backbone atoms. SS has three different major components: helix (H), beta (E) and coil (C) shown in Figure 1.

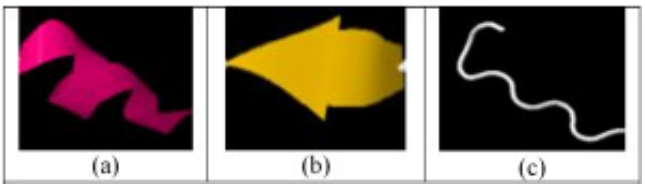

Fig. 1. Three major secondary structures: (a) helix (pink), (b) beta (yellow) and (c) coil (while).

Most of the SS predictors express imbalanced accuracies by claiming higher prediction performances in predicting *H* and *C*, and on the contrary having lower accuracy in *E* predictions. *E* component being in low count, a predictor may show good performance by over-predicting *H* and *C* and under predicting *E*. However, this under-prediction of *E* class can make such predictors biologically inapplicable. In this proposed work, we are motivated to develop a balanced SS predictor.

We developed a new statistical predictor of SS from protein sequence which constitutes 4 major steps: *First*, is to construct valid benchmark training and test datasets. We used one training dataset and two test datasets. The training dataset, namely T552, is collected from Protein Data Bank (PDB). T552 composed of 149,093 residues with 18.2%, 51.6% and 30.2% of *E*, *C* and *H* residues, respectively. It is to be noted that, secondary structure data set is naturally imbalanced. We developed two dataseta: CB513, N295 for the conducted empirical experiment.

Second, is to encode the protein sequence with feature sets that can appropriately distinguish the type of classes under consideration. We utilized 33 features per amino acid for this purpose.

Third, is to develop an effective prediction engine, where we have used binary SVMs coupled with genetic algorithm (GA). We have trained three binary Support Vector Machine (SVM)s: *i*) , *ii*) and *iii*) . Although SVM could have been used directly for three class classification, we rather choose to use three binary SVM classifiers, so that we can attain a balanced accuracy in all three classes. GA finds the real value parameter for each class as an additive factor for each class-probability given by three binary SVMs. We refer to our combined SVM predictor as *cSVM*. Our final predictor is a meta-predictor, named as MetaSSPred, combines the outputs of cSVM and SPINE X.

Four, is to evaluate the performance of the predictor. We performed the preliminary experiments on feature set and methodology selection. The results show that the set with 33 features performs the best among four feature sets with 29, 31, 33 and 51 features, respectively. The comparison of the results on both of the test datasets shows that three binary class classification (cSVM) classifier performs better than SVM that directly classifies three classes. The novel paradigm, MetaSSPred, significantly increases beta accuracy () for both the datasets. scores of MetaSSPred on CB471 and N295 were 71.7% and 74.4% respectively. These scores are 20.9% and 19.0% improvement over the scores given by SPINE X alone on CB471 and N295 datasets respectively. Standard deviations of the accuracies across three SS classes of MetaSSPred on CB471 and N295 datasets were 4.2% and 2.3% respectively. On the other hand, for SPINE X, these values are 12.9% and 10.9% respectively. These findings suggest that the proposed MetaSSPred is a well-balanced SS predictor. The software is available as a standalone software.

Publication

A balanced secondary structure predictor.
Nasrul Islam M, Iqbal S, Katebi AR, Tamjidul Hoque M
J Theor Biol. 2016 Jan 21