Review: Protein Fold Prediction

A. Dievsky

June 7, 2013

The prediction of a new protein's spacial structure forms an important task in proteomics. Naturally the first step is often to predict the secondary structure features (i.e., helices, sheets and coils, and buried and exposed amino acid residues). This step is called "fold recognition".

There are several techniques of the fold recognition. Firstly, the fold structure may be inferred from the known fold of a homologous protein. However, having a well-studied homologue is a rare case. Secondly, the fold structure can be imputed statistically from the various biochemical data such as the prevalence of certain amino acids in certain folds (this is called *threading*). The second class is more taxing, as most novel proteins do not have a readily available homologue with known structure, especially since the advent of mass-spectrometry.

The homology-based prediction methods enjoy high accuracy rate (homologous sequence are more than likely to be folded similarly), but suffer from low coverage (most sequences are not homologous to any known fold). Reversely, the threading methods offer high coverage (statistical data covering almost any sequence), but have low accuracy.

The article "A Hierarchical Approach to Protein Fold Recognition" by Mohammad and Nagarajaram describes a new method of predicting a protein fold based on statistical data. The approach employs the Support Vector Machine algorithm (SVM). While applying SVM to fold recognition isn't new *per se* (for example, see SVM-Fold method), the novelty is in the hierarchical nature of the prediction algorithm. First, only a broad structural class of a protein is predicted. Then, within the given class, the fold is predicted in more detail. The binary classifier of an ordinary SVM is extended to multiple classes using one-versus-one method.

The structural classes predicted by the algorithm are the four main structural classes from the SCOP database: all-α, all-β, α and β with parallel sheets (α/β) and α and β with antiparallel sheets ($\alpha + \beta$). The SVM uses amino acid singletons and pairs as features, the frequencies of these features being well-known for each structural class, structural state and solvent accessibility.

The discussed model achieves higher accuracy than its predecessors, namely 57% on a 79-folds dataset and 80% on a 711-folds dataset. Performance evaluation on Lindahl and Elofsson's dataset yields an accuracy score of 70%, which is almost twice higher than the previous record-setter, TAXFOLD (40%).

The authors inspected both levels of their classifier to find out which one is more responsible for the mispredictions. It turned out that most of the incorrectly predicted folds were attributed to a wrong structural class; when the fold was predicted within the correct class, the error rate diminished by about 30%. Thus, the main area of further improvement is identified as the first step of the method, i. e. the prediction of the structural class.

One should note that the paper is rather nonchalant with the details of the algorithm, redirecting the reader to previous works of the same authors at crucial points. This trait makes it fairly hard to understand how exactly the proposed method works, even if one knows the basic principles of the SVM. However, the main idea of the article indeed looks innovative and adequate, so the technical details can be overlooked in favour of a general comprehension.

References

[1] Mohammad, T. A. S. and Nagarajaram, H. A., *A Hierarchical Approach to Protein Fold Prediction,* Journal of Integrative Bioinformatics, 8(1):185, 2011

[2] Melvin, I. *et al.*, *SVM-Fold: a tool for discriminative multi-class protein fold and superfamily recognition,* BMC Bioinformatics, 8 Suppl 4:S2, 2007

[3] Schaeffer, R. D. and Daggett V., *Protein folds and protein folding,* Protein Eng. Des. Sel., 24(1-2):11-9, 2011