**Abstract**

Detecting subtle sequence similarities or remote homology is an important problem in computational biology. Discriminative methods such as support vector machine (SVM) provide state-of-the-art performance. The success of the SVM method depends on the feature set used to represent each protein. The features are very large and noise data may be introduced. In this paper, the authors focus on feature extraction for SVM protein classification. They use a latent semantic analysis (LSA) model which is an efficient technology of text categorization from natural language processing.

In document classification, a word-document matrix is constructed first. LSA is performed on the matrix to reduce its rank and produce condensed representations of the document vectors in the reduced space. Protein classification involves separating protein sequences into classes, whereas text-categorization assigns free-text documents to pre-defined categories. In order to apply a text categorization technique to protein classification, an analogy for words has to be identified. The authors studied various basic building blocks of proteins as the ‘words’ of ‘protein sequence language’ such as N-grams, patterns and motifs. N-grams are the set of all possible sub-sequence of amino acids of a fixed length N. They selected 8000 patterns as the characteristic words. The MEME/MAST system is used to discover motifs – the highly conserved regions of proteins. Each protein sequence is then considered to be a ‘document’ containing these ‘words’. Analogous to document classification, the ‘word-document’ matrix is constructed using the weight of each word in the documents and the LSA model is used to decrease the dimensions of the original protein vectors. The length of the resulting vector is much shorter than that of the original one resulting in noise removal and efficient description of the protein sequence. These condensed vectors are evaluated by SVM to train classifiers which are then used to classify the protein sequences.

The authors tested their model on the SCOP 1.53 database. They used ROC scores and the median rate of false positives to evaluate the experimental results. They compared the performance of three SVM-based methods after latent semantic analysis with PSI-BLAST, SVM-pairwise and SVM-LA. The performance was found to be comparable to that of complex kernel methods such as SVM-LA and better than that of sequence based methods such as PSI-BLAST and SVM-pairwise.

**Discussion**

The SVM–based methods discussed in class focused on finding useful representations of protein sequences using kernel functions. As the feature set is usually very large, such representations may suffer from noise and the peaking phenomenon in many machine learning methods. This paper presents a different approach for the representation of protein sequences by leveraging an efficient feature extraction technique from natural language processing – Latent Semantic Analysis. The latent semantic representation protein vectors are evaluated by SVM. The results show that the LSA model significantly improves performance of remote homology detection as compared to the basic SVM methods without LSA. The main bottleneck of the LSA method is the computational complexity involved in constructing the word-document matrix and reducing its rank. It roughly takes $O(nm/t)$, where $n$ is the number of training examples, $m$ is the total number of words and $t$ is the minimum of $m$ and $n$.  

---

**Additional Paper for “Machine Learning for Protein Classification”**

**Paper reference**

“Application of latent semantic analysis to protein remote homology detection”  