Abstract:
One key element in understanding the molecular machinery of the cell is to understand the structure and function of each protein encoded in the genome. A very successful means of inferring the structure or function of a previously un-annotated protein is via sequence homology with one or more proteins whose structure or function is already known. In this paper, a novel method for protein remote homology detection has been presented. The technologies of text categorization from natural language processing have been used in protein classification. Patterns are discovered by TEIRESIAS algorithm and can be viewed as the “words” of “protein sequence language”. The patterns are then filtered by an efficient feature selection algorithm called chi-square algorithm. Each protein sequence is mapped into a high dimensional vector by the occurrence times of the selected patterns. This presentation, combined with a discriminative classification algorithm known as the Support Vector Machine (SVM), provides a powerful means for protein remote homology detection. The method, called SVM-pattern, is tested on the SCOP database and compared with other state-of-the-art methods. The performance of SVM-pattern is better than that of BLAST method and comparable with other SVM-based methods such as SVM-k-spectrum and SVM-pairwise.

Keywords:
Protein; remote homology; text categorization; pattern

1. Introduction
A central problem in computational biology is the classification of proteins into functional and structural classes given their amino acid sequences. The three-dimensional structure of a protein largely determines its function in the cell. However, it is far easier to determine the primary sequence of a protein than to solve the three-dimensional structure. Through evolution, structure is more conserved than sequence, so that detecting even very subtle sequence similarities, or remote homology, is important for predicting the functions of proteins. While most methods can detect homology with a high level of similarity, remote homology is often difficult to separate from pairs of proteins that share similarities due to chance. Detecting homology in the so-called “twilight zone” remains challenging nowadays [1].

The major methods for homology detection can be split into three basic groups [2]: pair-wise sequence comparison algorithms, generative models for protein families and discriminative classifiers. Early methods looked for pair-wise similarities between proteins. Among such algorithms, the Smith-Waterman dynamic programming algorithm [3] is one of the most accurate methods, whereas heuristic algorithms such as BLAST [4] and FASTA [5] trade reduced accuracy for improved efficiency. The methods afterwards have obtained higher rate of accuracy by collecting statistical information from a set of similar sequences. PSI-BLAST [6] uses BLAST to iteratively build a probabilistic profile of a query sequence and obtains a more sensitive sequence comparison score. Generative models such as profile hidden Markov models (HMM) [7] use positive examples of a protein family, but they can be trained iteratively using both positively labeled and unlabeled examples by pulling in close homology and adding them to the positive set. Finally, the discriminative algorithms such as Support Vector Machine (SVM) [8] use both positive and negative examples and provide state-of-the-art performance with appropriate kernel. Many SVM-based methods have been proposed such as SVM-fisher [9], SVM-k-spectrum [10], SVM-pairwise [2], SVM-I-sites [11], SVM-LA and SVM-SW [1]. A comparison of SVM-based methods has been performed by Saigo et al. [12].

The mapping of primary sequences to structures and functions of proteins is conceptually similar to the mapping of words to meanings, thus the natural language processing techniques can be used to deal with the sequence, structure and function mapping problem. The similarity between biological sequence and natural language has recently been
2. Material and Method

Protein classification is the task of classifying the protein sequence into structure or function related class, whereas text categorization is the problem of assigning free text documents to predefined categories. The simple method for text categorization uses various kinds of classifiers based on characteristic words. Here, the patterns [18] were treated as the basic “words” of protein sequences and the support vector machine combining with the chi-square feature selection algorithm was used to detect the remote homology of protein sequence. The use of “biological words” is not novel. Many protein classification methods implicitly use various biological “words”, such as the k-spectrum [10], the motif [22], the I-sites [11] and the n-grams [16]. Using patterns as the protein secondary structure words Dong et al. [23] have successfully predicted the secondary structures of proteins.

2.1. Protein words

Many biological problems being addressed are analogous to the problem of processing an unknown natural language when the only available knowledge is the set of its most basic units. For natural languages, these most basic units can be either alphabet letters (e.g., Hebrew, Greek, English, and so forth) or syllabic signs (e.g., Chinese, Japanese, and so forth). In the absence of any language-specific information, one attempt to understand such a language would process in a bottom-up manner: from the most basic units, to the level of the vocabulary, then to the level of the syntax and semantics.

In proteins, the most basic units are the 20 natural amino acids. Unlike the natural language, the exact “words” of proteins are not clear. In the absence of any language information, one can think that a word is some consecutive basic units that usually occur together. Besides that, in evolution, amino acid substitution is a common phenomenon. Thus the protein words should be composed of amino acids and the wildcard (‘.’) that can be any of the twenty amino acids. Here are some examples: E.L.K, NGF, KI...L, Q..Y.A..L. The protein words used here are same as the patterns [18] in biological sequence analysis. Let $\Sigma$ be a alphabet, a pattern is a string in $\Sigma \cup \Sigma(\Sigma \cup \{\})^* \Sigma$, that is, a string on the alphabet $\Sigma \cup \{\}$ that starts and ends with a solid character (not wildcard). A pattern $P$ is also called a $<L, W>$ pattern if any substring of $P$ with length $W$ or more contains at least $L$ solid characters.

The TEIRESIAS [19] algorithm is used to extract words in protein sequences. TEIRESIAS is a deterministic algorithm that allows one to carry out pattern discovery in biological sequences and has been used to annotate proteins [24]. The algorithm carries out this task without need for a data model and generates all patterns appearing $K$ or more times in the input with the additional guarantee of pattern maximality in both composition and length.

In this study, the TEIRESIAS algorithm was executed on all the protein sequences (training and testing sequences) and totally 71009 patterns were extracted. The produced patterns contain too much redundant information and many machine learning methods can’t perform well in the high-dimensional feature space. It is highly desirable to reduce the native space without sacrificing categorization accuracy.

2.2. Chi-square feature selection

Most machine-learning algorithms do not scale well to high-dimensional feature spaces [25], and the TEIRESIAS algorithm generates too many protein “words” with redundancy. Thus, it is desirable to reduce the dimension of the feature space by removing non-informative or redundant features. A large number of feature selection methods have been developed for this task, including document frequency, information gain, mutual information, chi-square and term strength. The chi-square algorithm is
selected in this study because it is one of the most effective feature selection methods in document classification [20].

The chi-square algorithm measures the lack of independence between a feature $t$ and a classification category $c$ and can be compared to the chi-square distribution with one degree of freedom to judge extremeness. The chi-square value of feature $t$ relative to category $c$ is defined to be:

$$\chi^2(t,c) = \frac{N \times (A \times D - C \times B)^2}{(A + C) \times (B + D) \times (A + B) \times (C + D)}$$  \hspace{1cm} (1)

where $A$ is the number of times $t$ and $c$ co-occur, $B$ is the number of times $t$ occurs without $c$, $C$ is the number of times $c$ occurs without $t$, $D$ is the number of times neither $c$ nor $t$ occurs and $N$ is the total number of protein sequence.

The chi-square statistic has a natural value of zero if $t$ and $c$ are independent. The category-specific scores of each feature can be combined into two scores

$$\chi^2_{avg}(t) = \sum_{i=1}^{m} p_i(c_i) \chi^2(t,c_i)$$  \hspace{1cm} (2)

$$\chi^2_{max}(t) = \max_{i=1,m} \{\chi^2(t,c_i)\}$$  \hspace{1cm} (3)

In this paper, the maximum feature value is used, since its performance is better than the average value.

### 2.3. Remote homology detection by SVM

Support Vector Machine (SVM) is a class of supervised learning algorithms first introduced by Vapnik [8]. Given a set of labeled training vectors (positive and negative input examples), SVM can learn a linear decision boundary to discriminate between the two classes. The result is a linear classification rule that can be used to classify new test examples. SVM has exhibited excellent performance in practice and has strong theoretical foundation of statistical learning theory.

Suppose the training set $S$ consists of labeled input vectors $(x_i, y_i)$, $i = 1...m$, where $x_i \in R_n$ and $y_i \in \{\pm 1\}$. A linear classification rule can be specified:

$$f(x) = w \cdot x + b$$  \hspace{1cm} (4)

where $w \in R_n$ and $b \in R$. A test example $x$ is then classified as positive (negative) if $f(x) > 0$ ($f(x) < 0$). Such a classification rule corresponds to a hyper-plane decision boundary between positive and negative points with normal vector $w$ and bias term $b$. When the samples are not linearly separable, the kernel function is used to map the samples from the input space into the feature space in which an appropriate hyper-plane can be found.

In this study, the Gist SVM package implemented by Jaakkola et al. [9] is used for protein remote homology detection. Figure 1 has demonstrated the complete implementation of SVM-pattern method.

### 2.4. Data set

The standard evaluation data is same as the one used by Li et al. [2], which is taken from the Structural Classification of Proteins (SCOP) database [21] version 1.53. Sequences were selected using the ASTRAL database [26]. The data set contains 54 families and 4352 distinct sequences. Remote homology is simulated by holding out all members of a target 1.53 family from a given super-family. Positive training examples are chosen from the remaining families in the same super-family and negative test and training examples are chosen from outside the target family's fold. The held-out family members serve as positive test examples. Details of the data sets are available at http://www1.cs.columbia.edu/compbio/svm-pairwise/.

### 3. Result and Discussion

#### 3.1. Comparative methods

The primary experiment reported here compares the performance of four methods: BLAST [4],
SVM-k-spectrum [10], SVM-pairwise [2] and SVM-pattern. The accuracy of each method is evaluated by testing its ability to classify protein domains into super-families in the SCOP database version 1.53 [21]. The dataset has been described in section 2.4.

In SVM-pattern experiment, the TEIRESIAS algorithm is carried out on all the 4352 protein sequences and 71009 patterns are obtained in total. After chi-square algorithm, 10000 patterns are selected as the representative “words” of protein sequence. Each protein sequence is mapped into a high-dimensional vector by the occurrence times of each selected pattern.

BLAST [4] is a heuristic approximation of the dynamic programming sequence alignment method. The process of homology detection by BLAST is same as the one used by Grundy [27]. The similarity score of a testing sequence is the average score of the testing sequence against each of the sequence in training set.

The features used by the SVM-k-spectrum [10] method are the set of all possible subsequences of amino acids of a fixed length \(k\). The length of the feature vector increases exponentially with \(k\). Here, the value of \(k\) is taken as 3. It is sufficient to detect remote homology of protein sequence.

In SVM-pairwise method [2], the feature vector is a list of pairwise sequence similarity scores, computed with respect to all of the sequences in the training set.

All the SVM-based methods use the Gist SVM package [9] to train the classifier and produce the similarity scores of the testing samples.

3.2. Performance metrics

Each of the method produces a score for a testing protein sequence, which represents the similarity between the protein and the particular family. Two methods are used to evaluate the experiment results: the Receiver Operating Characteristic (ROC) scores [28] and the Median Rate of False Positives (M-RFP) scores [9]. The ROC score is the normalized area under a curve that plots true positives as a function of false positives for varying classification thresholds. A score of 1 indicates perfect separation of positive samples from negative samples, whereas a score of 0 denotes that none of the sequences selected by the algorithm is positive. The median RFP score is the fraction of negative test sequences that score as high or better than the median score of the positive sequences. Obviously, the small M-RFP value corresponds to good results. Hou et al. [11] have presented an efficient algorithm to compute these scores.

3.3. Experiment results

Table 1 summarizes the average ROC and M-RFP scores of the 54 SCOP families for the four methods. SVM-pattern method is significantly better than BLAST method and comparable with SVM-k-spectrum as well as SVM-pairwise method.

Figure 2 shows the detailed comparison of different methods. Each graph plots the total number of families for which a given method exceeds a score threshold. The left graph uses ROC scores and the right graph uses M-RFP.
scores. In each graph, a higher curve corresponds to more accurate homology detection performance. As seen in the figure, there is no significant different in the three SVM-based methods.

Table 1. The comparative performance of different methods

<table>
<thead>
<tr>
<th></th>
<th>BLAST</th>
<th>SVM-k-spectrum</th>
<th>SVM-pairwise</th>
<th>SVM-pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROC</td>
<td>0.718773</td>
<td>0.84639</td>
<td>0.825928</td>
<td>0.837715</td>
</tr>
<tr>
<td>M-RFP</td>
<td>0.085175</td>
<td>0.00757592</td>
<td>0.00754719</td>
<td>0.0213885</td>
</tr>
</tbody>
</table>

To further illustrate the difference of SVM-based method, the family-by-family comparison of two methods has been plot in figure 3. The axes are ROC scores achieved by the homology detection method labeled near the axes. Each point on the graph corresponds to one of the 54 SCOP families. When the families are in the right-bottom area, it means that the method labeled by x-axis outperforms the method labeled by y-axis on this family. One can see from the figure that the SVM-pattern method is a little worse than SVM-k-spectrum method, but is comparable with the SVM-pairwise method.

4. Conclusions

In this paper, a novel, simple and efficient method for protein remote homology detection has been presented. Its performance is comparable with other state-of-the-art methods. There are two main novelties in the method. The first novelty is that the technologies of text categorization from natural language processing have been used in protein classification. The secondary novelty is that a new building block of protein sequence, i.e. pattern, has been presented. Further work will aim at detail identification the syntaxes and semantics of the patterns to map out the relationship of protein sequence, structure and function.

Acknowledgements

The authors would like to thank Yong-sheng Yuan for his helpful discussions. Special thanks give to Xuan Liu for her comments on this work that significantly improve the presentation of the paper. Financial support was provided by the National Natural Science Foundation of China (60435020).

References


