BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 3 2006, pages 285–290
doi:10.1093/bioinformatics/bti801

Sequence analysis

Application of latent semantic analysis to protein remote homology detection

Qi-wen Dong*, Xiao-long Wang and Lei Lin

School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

Received on July 17, 2005; revised on November 6, 2005; accepted on November 24, 2005
Advance Access publication November 29 2005
Associate Editor: Charlie Hodgman

ABSTRACT

Motivation: Remote homology detection between protein sequences is a central problem in computational biology. The discriminative method such as the support vector machine (SVM) is one of the most effective methods. Many of the SVM-based methods focus on finding useful representations of protein sequence, using either explicit feature vector representations or kernel functions. Such representations may suffer from the peaking phenomenon in many machine-learning methods because the features are usually very large and noise data may be introduced. Based on these observations, this research focuses on feature extraction and efficient representation of protein vectors for SVM protein classification.

Results: In this study, a latent semantic analysis (LSA) model, which is an efficient feature extraction technique from natural language processing, has been introduced in protein remote homology detection. Several basic building blocks of protein sequences have been investigated as the 'words' of 'protein sequence language', including N-grams, patterns and motifs. Each protein sequence is taken as a 'document' that is composed of bags-of-word. The word-document matrix is constructed first. The LSA is performed on the matrix to produce the latent semantic representation vectors of protein sequences, leading to noise-removal and smart description of protein sequences. The latent semantic representation vectors are then evaluated by SVM. The method is tested on the SCOP 1.53 database. The results show that the LSA model significantly improves the performance of remote homology detection in comparison with the basic formalisms. Furthermore, the performance of this method is comparable with that of the complex kernel methods such as SVM-LA and better than that of other sequence-based methods such as PSI-BLAST and SVM-pairwise.

Availability: The source codes are freely available at http://www.insun.hit.edu.cn/news/view.asp?id=413 or upon request from the authors.

Contact: qwdong@insun.hit.edu.cn

INTRODUCTION

A central problem in computational biology is the classification of proteins into functional and structural classes given their amino acid sequences. Through evolution, structure is more conserved than sequence. Therefore, detecting very subtle sequence similarities, or remote homology, is important for predicting the functions of proteins. Most methods can detect homology with a high level of similarity, while remote homology is often difficult to be separated from pairs of proteins that share similarities owing to chance. Detecting homology in the so-called 'twilight zone' remains challenging nowadays (Saigo *et al*., 2004).

The major methods for homology detection can be split into three basic groups (Li and Noble, 2003): pairwise sequence comparison algorithms, generative models for protein families and discriminative classifiers. Early methods looked for pairwise similarities between proteins. Among those algorithms, the Smith–Waterman dynamic programming algorithm (Smith and Waterman, 1981) is one of the most accurate methods, whereas heuristic algorithms such as BLAST (Altschul *et al*., 1990) and FASTA (Pearson, 1990) 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 (Altschul *et al*., 1997) used BLAST to iteratively build a probabilistic profile of a query sequence and obtained a more sensitive sequence comparison score. Generative models such as profile hidden Markov models (HMM) (Karplus *et al*., 1998) used positive examples of a protein family, which can be trained iteratively using both positively labeled and unlabeled examples by pulling in close homology and adding them to the positive set (Qian and Goldstein, 2004). Finally, the discriminative algorithms such as support vector machine (SVM) (Vapnik, 1998) used both positive and negative examples and provided state-of-the-art performance with appropriate kernel. Many SVM-based methods have been proposed such as SVM-fisher (Jaakkola *et al*., 2000), SVM-k-spectrum (Leslie *et al*., 2002), Mismatch-SVM (Leslie *et al*., 2004), SVM-pairwise (Li and Noble, 2003), SVM-I-sites (Hou *et al*., 2003), SVM-LA and SVM-SW (Saigo *et al*., 2004). A comparison of SVM-based methods has been performed by Saigo *et al*. (2002).

The success of a SVM classification method depends on the choice of the feature set to describe each protein. Most of these research efforts focus on finding useful representations of protein sequence data for SVM training by using either explicit feature vector representations or kernel functions. The features are usually very large and noise data may be introduced. In contrast, this research focuses on the feature extraction for SVM protein classification. Especially, a latent semantic analysis (LSA) model from natural language processing (Bellegarda, 2000) has been introduced to condense the original protein vectors. The length of the resulting vector is much shorter than that of the original vector leading to noise-removal and efficient description of the protein sequence.

As a proven method in the case of natural language processing, LSA has been used to generate summaries, compare documents and

*To whom correspondence should be addressed.

retrieve further information (Bellegarda, 2000). Recently, LSA was also introduced in computational biology and used to predict the secondary structure of protein (Ganapathiraju *et al*., 2004). Furthermore, the similarity between biological sequence and natural language has recently attracted much attention. Many methods of natural language processing have been applied to biological sequences. The N-grams of whole genome protein sequences have been analyzed and some statistical features have been extracted (Ganpathiraju *et al*., 2002). The probabilistic models from speech recognition have been employed to enhance the protein domain discovery (Coin *et al*., 2003). Protein classification based on text document classification techniques has provided state-of-the-art performance on GPCR classification (Cheng *et al*., 2005). The protein sequence language has been discussed extensively by Ganapathiraju *et al*. (2005).

In this paper, the technologies of text categorization from natural language processing have been used in protein classification. A method by combining LSA with SVM has been presented for protein remote homology detection. Various 'words' of 'protein sequence language' have been investigated, including N-grams (Leslie *et al*., 2002), patterns (Dong *et al*., 2005) and motifs (Ben-Hur and Brutlag, 2003). Experimental results showed that the use of LSA technology significantly improves the performance of protein remote homology detection.

SYSTEMS AND METHODS

Method overview

Protein classification is the task to separate the protein sequences into structure- or function-related classes, whereas text categorization is the problem of assigning free text documents to predefined categories. In order to apply text categorization techniques to protein sequences, first a suitable analogy for words has to be identified. Here, three basic building blocks including N-grams (Leslie *et al*., 2002), patterns (Dong *et al*., 2005) and motifs (Ben-Hur and Brutlag, 2003) have been introduced as the 'words' of proteins. The N-grams are the set of all possible subsequences of amino acids of a fixed length N. In this study, the value of N is taken as 3, so the total words of protein sequences are 8000 (20^3). The patterns (Pisanti *et al*., 2002) denote strings on the alphabet $\Sigma \cup \{'.'\}$, where Σ is the set of the 20 amino acids and $\{'.'\}$ can be any of the amino acids. The TEIRESIAS (Rigoutsos and Floratos, 1998) algorithm is used to extract patterns in protein sequences with parameters $L = 3$, $W = 6$, $K = 10$ and totally 71 009 patterns are extracted. Since many machine learning methods cannot perform well in the high-dimensional feature space, it is highly desirable to reduce the native space by removing non-informative or redundant patterns. After an effective feature selection (χ^2 selection), 8000 patterns are selected as the characteristic words. Motifs denote the limited, highly conserved regions of proteins. The MEME/MAST system version 3.0 (Bailey and Elkan, 1994) is used to discover motifs and search databases. Since motifs only exist in related protein sequences, the training sequences of the same superfamily are used to generate motifs. Totally, there are 3231 motifs extracted. For a detailed description of these basic building blocks, please refer to the supplement notes.

In order to apply LSA to protein remote homology detection, each protein sequence that belongs to a particular class is treated as a 'document' that is composed of bags-of-X, where X can be any basic building blocks of protein sequences. The word-document matrix needs to be constructed by collecting the weight of each word in the documents. Figure 1 presents an example of such matrices. Singular value decomposition (SVD) is performed on the word-document matrix to remove the noise from the data and to decrease the dimensions of the protein vectors. The latent semantic representation vectors

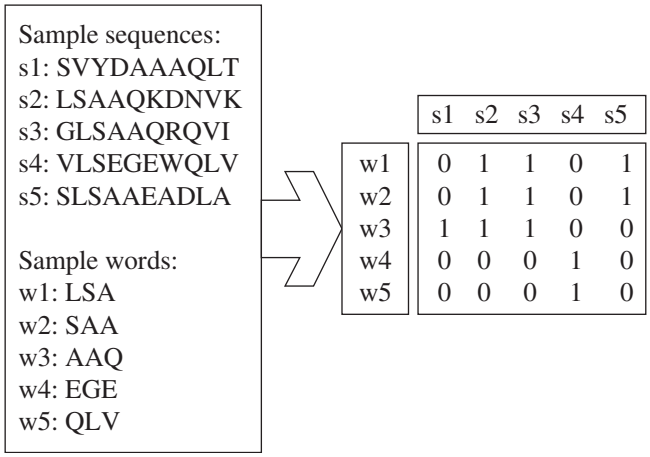

Fig. 1. Sample construction of the word-document matrix with N-grams as the words. The cell entries are the times of occurrence of a word (rows) in a document (columns).

are evaluated by support vector machine to train classifiers which are then used to classify the test protein sequences.

In this study, the Gist SVM package implemented by Jaakkola *et al*. (2000) is applied for protein remote homology detection. The parameters of SVM are used by default of the Gist package except that the kernel function i.e. the radius basis function (RBF) kernel. Figure 2 illustrates the implementation of the method.

Latent semantic analysis

LSA is a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text (Landauer *et al*., 1998). Here, we briefly describe the basic process of LSA.

The starting point of LSA is the construction of a word-document matrix W of co-occurrences between words and documents. The elements of W can be taken as the number of times each word appears in each document, thus the dimension of W is $M \times N$, where M is the total number of words and N is the number of given documents. To compensate for the differences in document lengths and overall counts of different words in the document collection, each word count can be normalized (Landauer *et al*., 1998).

In the word-document matrix W, each document is expressed as a column vector. However, this representation does not recognize synonymous or related words and the dimensions are too large. In the specific application, singular value decomposition is performed on the word-document matrix. Let K be the total ranks of W, W can be decomposed into three matrices:

$$W = USV^{\mathrm{T}} \tag{1}$$

where U is left singular matrix with dimensions $(M \times K)$, V is right singular matrix with dimensions $(N \times K)$ and S is $(K \times K)$ diagonal matrix of singular values $s_1 \geq s_2 \geq \cdots s_K > 0$. One can reduce the dimensions of the solution simply by deleting the smaller singular values in the diagonal matrix. The corresponding columns of matrix U (rows of matrix V) are also ignored. In practice only the top R ($R \ll \mathrm{Min}\,(M, N)$) dimensions for which the elements in S are greater than a threshold are considered for further processing. Thus, the dimensions of matrices U, S and V are reduced to $M \times R$, $R \times R$ and $N \times R$, leading to data compression and noise removal. Values of R in the range [200, 300] are typically used for information retrieval. In the present context, the best results are achieved when R takes the value of $\sim$300.

By SVD, the column vectors of the word-document matrix W are projected onto the orthonormal basis formed by the row column vectors of the

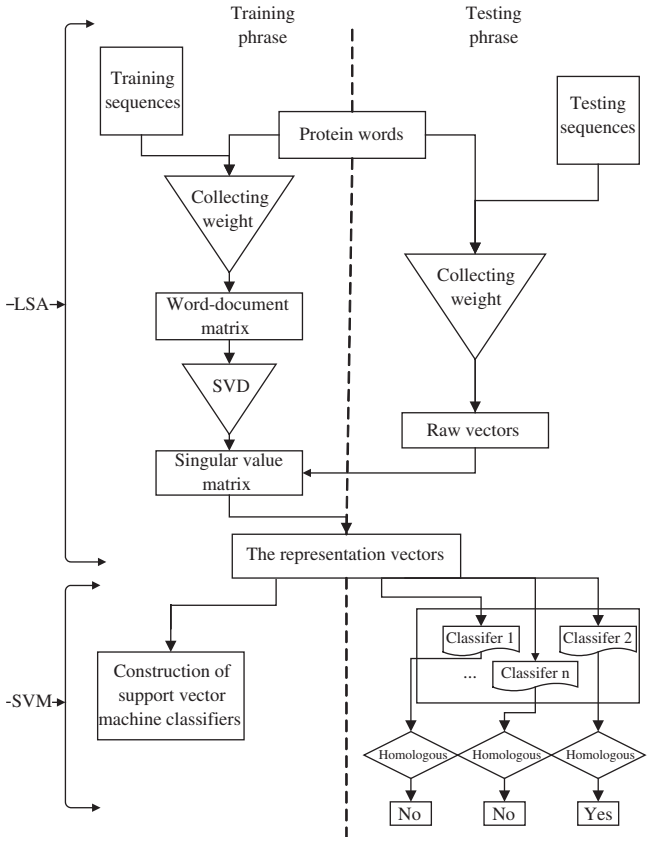

Fig. 2. Overview of LSA-based SVM for protein classification. The word-document matrix is constructed by the context of protein sequences. The latent semantic analysis is then performed on the matrix to produce the latent semantic representation vectors of protein sequences. The support vector machine is used to evaluate the protein vectors. Such systems can use any building blocks of proteins as the protein words.

left singular matrix U. The coordinates of the vectors are given by the columns of SV^{T}. This in turn means that the column vectors Sv_j^{T} or, equivalently the row vector v_jS, characterizes the position of document d_j in the R dimensions space. Each of the vector v_jS is referred to a document vector, uniquely associated with the document in the training set.

For a new document that is not in the training set, it is required to add the unseen document to the original training set and the latent semantic analysis model be recomputed. However, SVD is a computationally expensive process, performing SVD every time for a new test document is not suitable. From the mathematical properties of the matrices U, S and V, the new vector t can be approximated as

$$t = dU, \tag{2}$$

where d is the raw vector of the new document, which is similar to the columns of the matrix W.

Data set and performance metrics

The standard evaluation data are the same as the one used by Li *et al*. (2003), which is taken from the Structural Classification of Proteins (SCOP) database (Andreeva *et al*., 2004) version 1.53. Sequences are selected from the ASTRAL database (Chandonia *et al*., 2004). The dataset 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 superfamily. Positive training examples are chosen from the remaining families in the same superfamily and negative test and training examples are chosen from outside the fold of the target family. The held-out family members serve as positive test examples. This process is iterated until each family has been tested. Details of the datasets are available at http://www1.cs.columbia.edu/compbio/svm-pairwise/

Two methods are used to evaluate the experimental results: the receiver operating characteristic (ROC) scores (Gribskov and Robinson, 1996) and the median rate of false positives (M-RFP) scores (Jaakkola *et al*., 2000). An ROC score is the normalized area under a curve that is plotted with 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 smaller the M-RFP is, the better the results are.

Setup of competing method

Through the experiments reported here, the performances of the following methods are compared: PSI-BLAST, SVM-pairwise, SVM-LA, three SVM-based methods (including SVM-Ngram, SVM-Pattern and SVM-Motif) and three SVM-based methods after latent semantic analysis (including SVM-Ngram-LSA, SVM-Pattern-LSA, SVM-Motif-LSA). The setup procedures of these methods are briefly described as follows.

The PSI-BLAST is probably the most widely applied protein homology detection algorithm that only requires a single sequence as input. But for better performance, multiple sequences are input to PSI-BLAST. First, a random positive training sequence is selected as the initial query. The complete positive training set is then aligned by the CLUSTALW method (Thompson *et al*., 1994). Using the query sequence and the alignment as inputs, PSI-BLAST is run with the test set as a database. The resulting *E*-values are used to rank the test set.

For the SVM-based method, the key step is to express a protein sequence as a vector or the calculation of kernels. In the SVM-pairwise method (Li and Noble, 2003), the feature vector is a list of pairwise sequence similarity scores, computed with respect to all of the sequences in the training set. In the SVM-LA method (Saigo *et al*., 2004), the kernel is calculated by summing up scores obtained from the local alignments with gaps of the sequences. Such kernel may not be a positive definite kernel and the authors provided two solutions for this problem. Due to its performance and simplicity, we have implemented one of the methods, namely, the LA-ekm kernel. The parameters of LA-ekm kernel take the optimal values provided by the authors ($\beta = 0.5$, $d = -11$, $e = -1$). For the SVM method based on three basic words, the length of the feature vector is equal to the number of each type of words. A protein sequence is mapped to a high-dimensional vector by the frequency of occurrence of each word. The protein vectors are then input into SVM to train the classifiers and classify the test protein sequences. Such representation is also used in related work (Ben-Hur and Brutlag, 2003; Dong *et al*., 2005; Leslie *et al*., 2002). For the LSA-based method, the word-document matrix is built by collecting the weight of each word in the documents. LSA is then performed on this matrix to produce the latent semantic representative vectors of protein sequences. The complete pipeline is shown in Figure 2.

RESULTS AND DISCUSSION

Table 1 summarizes the performance of the various methods in terms of average ROC and M-RFP scores over all 54 families tested. The distributions of ROC and M-RFP scores are plotted in Figure 3. In each graph, a higher curve corresponds to more accurate homology detection performance. As seen in the figure, the PSI-BLAST

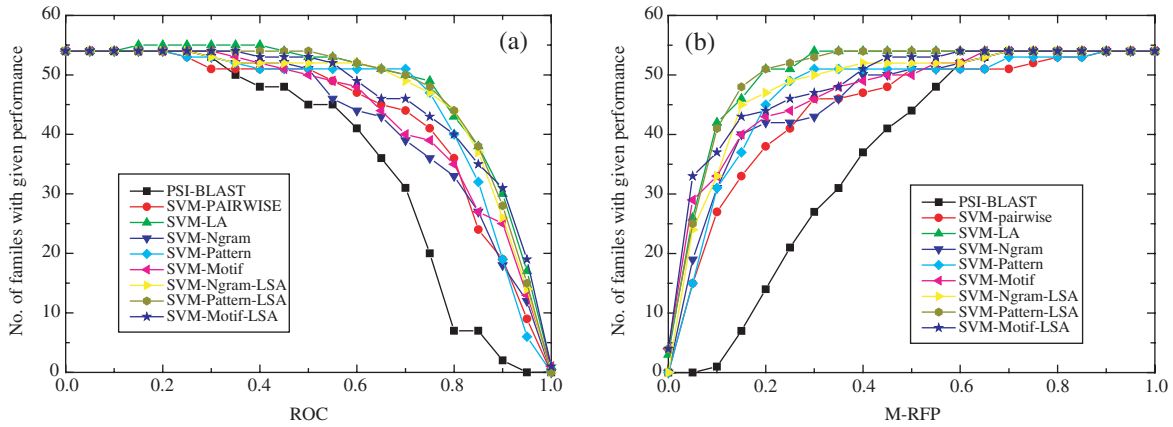

Fig. 3. Relative performance of homology detection methods. The graph plots the total number of families for which the method exceeds a given performance. Each series corresponds to one of the homology detection methods described in the text. The left part (**a**) uses the ROC scores and the right part (**b**) uses the M-RFP scores.

Table 1. Average ROC and M-RFP scores over 54 families for different methods

Methods	Mean ROC	Mean M-RFP
PSI-BLAST	0.675393	0.325322
SVM-pairwise	0.825928	0.1173329
SVM-LA	0.887124	0.0653927
SVM-Ngram	0.791415	0.144053
SVM-Pattern	0.835387	0.134893
SVM-motif	0.81356	0.124572
SVM-Ngram-LSA	0.859484	0.101688
SVM-Pattern-LSA	0.878926	0.070287
SVM-Motif-LSA	0.859193	0.0995269

method achieves the lowest performance. The accuracies of the SVM methods based on the basic words are lower than that of SVM-pairwise except for the pattern-based SVM method. When the LSA model is used, all the SVM methods based on the three basic words get higher accuracies. The performance of LAS model is comparable with that of the SVM-LA method and better than that of the SVM-pairwise method. The SVM-pairwise is one of the state-of-the-art methods and outperforms many other methods such as FPS (Bailey and Grundy, 1999), SAM (Krogh *et al*., 1994) and SVM-Fisher (Jaakkola *et al*., 2000), so the LSA model is an efficient method for remote homology detection.

When the three basic words are considered, one can find that the method based on patterns performs best whether the LSA model is used or not. The reason may be that there are wildcard in patterns. So patterns can match the protein sequences easily and describe the components of protein sequences effectively.

To present a better illustration of the difference between the methods with LSA and those without LSA, the family-by-family comparison of the ROC scores between the two methods has been plotted in Figure 4. Each point on the graph corresponds to one of the 54 SCOP families. When the families are in the left-upper area, it means that the method labeled by *y*-axis outperforms the method labeled by *x*-axis on this family. Obviously, all the methods with LSA can significantly outperform the methods without LSA.

The homology between the training samples and the test samples is an important factor that influences the performance of various methods. The contribution of homology to various methods is evaluated at the family, the superfamily and the fold level respectively. At the family level, the members of the target family are divided into two parts, one for positive training, and the other for positive test. At the superfamily level, the positive training samples are taken from the same superfamily of the target family, but the members from the family itself are excluded. At the fold level, the positive training samples are taken from the same fold of the target family, but the members from the superfamily of the target family are excluded. The negative training and test samples are same as those of previous experiments. Since many of the families contain unsuitable positive samples, only one of the families (SCOP ID: 2.1.1.4) is selected as the target family. The number of samples is listed in Table 2 and the results of various methods are listed in Table 3. At the family level, all the SVM-based methods perform equally well to PSI-BLAST. While at the superfamily level and the fold level, the improvement of SVM-based methods in comparison with PSI-BLAST is significant. So the discriminative methods are more powerful than PSI-BLAST for the detection of remote homology.

Computational efficiency is an important factor for any homology detection algorithm. In this regard, the LSA approaches are better than SVM-pairwise and SVM-LA but a little worse than the methods without LSA and PSI-BLAST. Any SVM-based method includes a vectorization step and an optimization step. The vectorization step of SVM-pairwise takes a running time of $O(n^2l^2)$, where n is the number of training examples and l is the length of the longest training sequence. The time complexity of calculation of LA-ekm kernel matrix is same as that of SVM- pairwise (Saigo *et al*., 2004). The time complexity of the vectorization step of the method without LSA is $O(nml)$, where m is the total number of words. The main bottleneck of the LSA method is the additional SVD process, which roughly takes $O(nmt)$, where t is the minimum of n and m. The optimization step of SVM-based method takes $O(n^2p)$ time, where p is the length of the latent semantic representation vector. In SVM-pairwise, p is equal to n, yielding a total time of $O(n^3)$. In the method without LSA, p is equal to m. While in the LSA method, p is equal to R. Since $R \ll \mathrm{Min}(n, m)$, the SVM

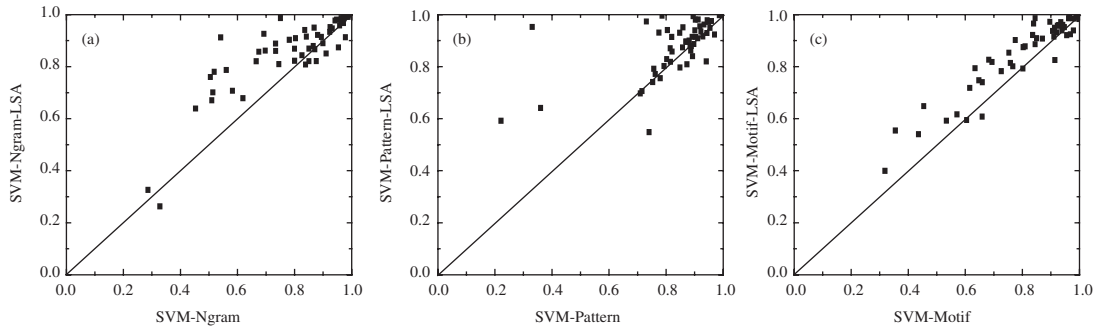

Fig. 4. Family by family comparison of the methods with LSA and those without LAS. The coordinates of each point in the plot are the ROC scores for one SCOP family, obtained by the two methods labeled near the axis. Figure (a), (b) and (c) are based on N-grams, patterns and motifs respectively.

Table 2. The numbers of samples at different homology level

	Positive train	Positive test	Negative train	Negative test
Family	20	13	3033	1137
Superfamily	88	33	3033	1137
Fold	61	33	3033	1137

The numbers of samples of the target family (2.1.1.4) at different homology level is listed. The selection of the samples for training and test is described in the main text.

Table 3. Comparative results of various methods at the family, the superfamily and the fold level

	Family		Superfamily		Fold	
	ROC	M-RFP	ROC	M-RFP	ROC	M-RFP
PSI-BLAST	0.9874	0.00082	0.8424	0.0219	0.6568	0.6525
SVM-LA	0.9986	0.00084	0.9857	0.0042	0.8942	0.0937
SVM-Ngram	0.8829	0.03078	0.8712	0.0386	0.7875	0.1143
SVM-Pattern	0.9983	0.00096	0.9759	0.0007	0.8639	0.0836
SVM-motif	0.9998	0.00073	0.9885	0.0008	0.8503	0.0993
SVM-Ngram-LSA	0.8929	0.05628	0.8992	0.0659	0.8455	0.1116
SVM-Pattern-LSA	0.9964	0.00098	0.9925	0.0017	0.9127	0.0674
SVM-Motif-LSA	0.9995	0.00087	0.9867	0.0035	0.9084	0.0721

The family (2.1.1.4) is the target family.

optimization step of LSA method is much faster than those of the other two methods. The time complexity of running PSI-BLAST is $O(nN)$, where N is the size of the database. In the current situation, N is approximately equal to nl.

The analysis presented here is based on sequences alone without using any evolutionary or structural information. Three basic building blocks of protein sequences are investigated: the N-grams, the patterns and the motifs. All of them show improved performance when the LSA model is used. Obviously, the structural or evolutionary information can further improve the performance of remote homology detection. Han *et al*. (2005) used profile–profile alignment and SVM for fold recognition. Hou *et al*. (2004) used local sequence–structure correlations for remote homology detection. Multiple profiles have been used for effective detection of remote homologues (Anand *et al*., 2005). Such evolutionary or structural information can also be used in LSA model, so long as the structural or functional building blocks of proteins are extracted. However, the identification of functional equivalents of 'words' in protein sequences is the major hurdle in the use of natural language techniques for a variety of computational biology problems (Ganapathiraju *et al*., 2005). In essence, the method presented here provides a fertile ground for further experimentation with dictionaries that can be constructed using different properties of the amino acids and proteins.

CONCLUSION

In this paper, the LSA model from natural language processing is successfully used in protein remote homology detection and improved performances have been acquired in comparison with the basic formalisms. Each document is represented as a linear combination of hidden abstract concepts, which arise automatically from the SVD mechanism. LSA defines a transformation between high-dimensional discrete entities (the vocabulary) and a low-dimensional continuous vector space S, the R-dimensional space spanned by the Us, leading to noise removal and efficient representation of the protein sequence. As a result, the LSA model achieves better performance than the methods without LSA.

Successful application of LSA to protein remote homology detection is of great significance. There are many problems in the biology domain that can be formulated as a classification task. Most of them, like fold prediction, tertiary structure and functional properties of proteins, are considered to be challenging problems. Thus, these important classification tasks are potential areas for applications of human language technologies in modern proteomics.

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).

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.

Altschul,S.F. *et al*. (1997) Gapped Blast and Psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

Anand,B. *et al*. (2005) Use of multiple profiles corresponding to a sequence alignment enables effective detection of remote homologues. *Bioinformatics*, **21**, 2821–2826.

Andreeva,A. *et al*. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res*., **32**, D226–D229.

Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, Menlo Park, California. pp. 28–36.

Bailey,T.L. and Grundy,W.N. (1999) Classifying proteins by family using the product of correlated *p*-values. In *Proceedings of the Third International Conference on Computational Molecular Biology (RECOMB99)*, Lyon, France, pp. 10–14.

Bellegarda,J. (2000) Exploiting latent semantic information in statistical language modeling. *Proc. IEEE*, **88**, 1279–1296.

Ben-Hur,A. and Brutlag,D. (2003) Remote homology detection: a motif based approach. *Bioinformatics*, **19**(Suppl 1), i26–i33.

Chandonia,J.M. *et al*. (2004) The ASTRAL Compendium in 2004. *Nucleic Acids Res*., **32**, 189–192.

Cheng,B.Y. *et al*. (2005) Protein classification based on text document classification techniques. *Proteins*, **58**, 955–970.

Coin,L. *et al*. (2003) Enhanced protein domain discovery by using language modeling techniques from speech recognition. *Proc. Natl Acad. Sci. USA*, **100**, 4516–4520.

Dong,Q.W. *et al*. (2005) A pattern-based SVM for protein remote homology detection. In *The Fourth International Conference on Machine Learning and Cybernetics*, GuangZhou, China, pp. 3363–3368.

Ganapathiraju,M. *et al*. (2005) Computational Biology and Language. *Ambient Intelligence for Scientific Discovery, LNAI*, **3345**, 25–47.

Ganapathiraju,M. *et al*. (2004) Characterization of protein secondary structure, Application of latent semantic analysis using different vocabularies. *IEEE Signal Processing Magazine*, **21**, 78–87.

Ganapathiraju,M. *et al*. (2002) Comparative N-gram analysis of whole-genome protein sequences. In *Proceedings of the Human Language Technologies Conference,* San Diego.

Gribskov,M. and Robinson,N.L. (1996) use of receiver operating characteristic(ROC) analysis to evaluate sequence matching. *Comput. Chem*., **20**, 25–33.

Han,S. *et al*. (2005) Fold recognition by combining profile–profile alignment and support vector machine. *Bioinformatics*, **21**, 2667–2673.

Hou,Y. *et al*. (2004) Remote homolog detection using local sequence-structure correlations. *Proteins*, **57**, 518–530.

Hou,Y. *et al*. (2003) Efficient remote homology detection using local structure. *Bioinformatics*, **19**, 2294–2301.

Jaakkola,T. *et al*. (2000) A discriminative framework for detecting remote protein homologies. *J. Comput. Biol*., **7**, 95–114.

Karplus,K. *et al*. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846–856.

Krogh,A. *et al*. (1994) Hidden Markov models in computational biology: applications to protein modeling. *J. Mol. Biol*., **235**, 1501–1531.

Landauer,T.K. *et al*. (1998) Introduction to latent semantic analysis. *Discourse Process*, **25**, 259–284.

Leslie,C. *et al*. (2004) Mismatch string kernels for discriminative protein classification. *Bioinformatics*, **20**, 467–476.

Leslie,C., Eskin,E. and Noble,W.S. (2002) The spectrum kernel: a string kernel for SVM protein classification. In *Proceedings of the Pacific Symposium on Biocomputing*, Hawaii, USA, pp. 564–575.

Li,L. and Noble,W.S. (2003) Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. *J. Comput. Boil*., **10**, 857–868.

Pearson,W.R. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol*., **183**, 63–98.

Pisanti,N., Crochemore,M., Grossi,R. and Sagot,M.F. (2002) A basis for repeated motifs in pattern discovery and text mining. *IGM 2002-10, Juillet*.

Qian,B. and Goldstein,R.A. (2004) Performance of an iterated T-HMM for homology detection. *Bioinformatics*, **20**, 2175–2180.

Rigoutsos,I. and Floratos,A. (1998) Combinatorial pattern discovery in biological sequences: the TEIRESIAS algorithm. *Bioinformatics*, **14**, 55–67.

Saigo,H. *et al*. (2004) Protein homology detection using string alignment kernels. *Bioinformatics*, **20**, 1682–1689.

Saigo,H. *et al*. (2002) Comparison of SVM-based methods for remote homology detection. *Genome Inform*., **13**, 396–397.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol*., **147**, 195–197.

Thompson,J.D. *et al*. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*., **22**, 4673–4680.

Vapnik,V.N. (1998) *Statistical Learning Theory*. Wiley, New York.