Protein Remote Homology Detection by Combining Profile-based Protein Representation with Local Alignment Kernel

Bin Liu[1,2,3], Xiaolong Wang[1,2], Ruifeng Xu[1,2], and Buzhou Tang[1,4]
[1] School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong, China;
[2] Key Laboratory of Network Oriented Intelligent Computation, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong, China;
[3] Shanghai Key Laboratory of Intelligent Information Processing, Shanghai, China
[4] School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, Texas, USA
Email: bliu@insun.hit.edu.cn

***Abstract*—Protein remote homology detection has attracted a great deal of interest as it is one of the most important problems in bioinformatics. Profile-based methods recently achieve the state-of-the-art performance. A key step to improve the performance of these methods is to find a suitable approach to use the evolutionary information in the profiles. In this study, we propose the profile-based protein representation to extract the evolutionary information from frequency profiles. In this approach, the frequency profiles calculated from the multiple sequence alignments outputted by PSI-BLAST are converted into several profile-based proteins and then the local alignment kernel (LA) is combined with these profile-based proteins for the prediction. Our experiments on a well-known benchmark show that the proposed approach can significantly improve the predictive performance.**

***Index Terms*—Protein remote homology, Support Vector Machine, profile-based proteins**

I. Introduction

By March 2013, 89003 experimentally determined protein structures were deposited in the Protein Protein Data Bank (PDB) [1], However, this number appears relatively small compared with the 539616 protein sequences held in the UniProtKB/Swiss-Prot database [2]. This vast amount of protein sequences need to be classified into structural and functional classes by means of homologies. Therefore, accurate computational methods that can automatically detect the protein remote homologies are needed. Unfortunately, protein remote homology detection is still a challenging problem in bioinformatics.

Early method challenge this problem by using pairwise comparison algorithms, such as BLAST [3] and Smith-Waterman local alignment algorithm [4]. However, in many cases these methods fail to detect remote homologies due to the low sequence similarities. Later methods challenge this problem by employing the generative models, which induce a probability distribution over the protein family and try to generate the unknown proteins as new member of the family from the stochastic model. For example, hidden Markov model (HMM) [5] can be trained iteratively in a semi-supervised manner, which uses both positively labeled and unlabeled samples of a particular family by pulling in close homology and adding them to the positive set [6].

Recently, discriminative methods, such as support vector machine (SVM) [7], challenge this problem with increasing success, which focus on the differences between protein families. The SVM-based methods lean a combination of the features that can discriminate the protein families. The main difference among these methods is kernel function, which computes the inner product between two samples in the feature space. The most straightforward approaches for generating the kernels are based on the features extracted from protein sequences. SVM-Ngram [8], SVM-pairwise [9] and SVM-LA [10] are three of the most successful sequence-based kernels. SVM-Ngram [8] is based on the feature space consisting all short subsequence of length *N*. In SVM-pairwise [9], a protein sequence is represented as a vector of pairwise similarities to all protein sequences in the training set, and then inner product between these vector-space representations is taken as the kernel. SVM-LA [10] measures the similarity between a pair of proteins by taking all the optimal local alignment scores with gaps between all possible subsequences into account. Besides these kernels, several other sequence-based kernels are also proposed, such as the motifs [11]-[13], mismatch [14], SVM-I-sites [15], SVM-n-peptide [16], N-gram [17], Patterns [18], SVM-BALSA [19], etc. The profile-based kernels further improve the performance by employing the evolutional information extracted from the profiles. For example, Top-*n*-grams [20] extract the profile-based patterns by considering the most frequent elements in the profiles. Profile kernel [21] extracts the short substrings according to the profile-based ungapped alignment scores. SW-PSSM [22] employs the profile-to-

Manuscript received June 12, 2013; revised August 28, 2013.

doi: 10.12720/jomb.3.1.17-22

profile scoring schemes for measuring the similarity between pairs of proteins. The recently proposed SVM-ACC method [23] treats the protein sequence as a time sequence and applies the auto-cross covariance (ACC) transformation to capture the correlation between any two properties in the profiles. Some profile-based methods improve the predictive performance by developing more sensitive profiles. HHsearch method [24] is based on a novel profile based on hidden Markov models. COMPASS [25] generates numerical profiles, constructs optimal profile-profile alignments and estimates the statistical significance of the corresponding alignment scores. Some web servers of profile-based algorithms are available online, including Bioshell [26], FORTE [27], COMA [28], PHYRE [29], GenThreader [30], and webPRC [31].

Some other features and techniques have been applied to this field in order to further improve the predictive performance. VBKC [32] uses a single multi-class kernel machine that combines kernels based on different feature space. Our recently proposed SVM-PDT [33] combines amino acid physicochemical properties and the profile features by physicochemical distance transformation (PDT), which is able to include the local sequence-order information of the entire protein sequences. The natural language processing techniques have been applied to this field. These methods are based on the similarities between protein sequences and natural languages. For example, our prior work shows that the performance of building-block-based methods can be improved by using the latent semantic analysis (LSA) [8]. PROTEMBED [34] learns an embedding of protein sequences into a low-dimensional semantic space for protein remote homology detection.

As introduced above, most of the top performing methods are based on features extracted from profiles, because a profile is a richer encoding of protein sequence than the individual sequence. The key step to improve the performance of these methods is to find a suitable approach to extract the evolutionary information from the profiles. In this article, we propose a protein-based protein representation to extract the evolutionary information from the frequency profiles. The frequency profiles are calculated from the multiple sequence alignments outputted by PSI-BLAST [35] and converted into a series of profile-based proteins. The Local Alignment kernel (SVM-LA [10]) is performed on these profile-based proteins. Testing on the SCOP 1.53 benchmark, we show that the proposed profile-based protein representation approach can obviously improve the performance of the Local Alignment kernel.

II. Methods

A. Dataset Description

A common benchmark [9] was used to evaluate the performance of our method for protein remote homology detection, which is available at http://noble.gs.washington.edu/proj/svm-pairwise/. This benchmark has been used by many studies of remote homology detection methods [8], [10], [36], which can provide good comparability with previous methods. The benchmark contains 54 families and 4352 proteins selected from SCOP version 1.53. These proteins are extracted from the Astral database [37] and include no pair with a sequence similarity higher than an E-value of 10^{-25}. For each family, the proteins within the family are taken as positive test samples, and the proteins outside the family but within the same superfamily are taken as positive training samples. Negative samples are selected from outside of the superfamily and are separated into training and test sets.

B. Frequency Profiles

The frequency profile ***M*** of protein *p* with *L* amino acids can be represented as:

$$M = \begin{bmatrix} m_{1,1} & m_{1,2} & \cdots & m_{1,L} \\ m_{2,1} & m_{2,2} & \cdots & m_{2,L} \\ \vdots & \vdots & \vdots & \vdots \\ m_{20,1} & m_{20,2} & \cdots & m_{20,L} \end{bmatrix} \tag{1}$$

where 20 is the total number of standard amino acids; $m_{i,j}$ ($0 \le m_{i,j} \le 1$) is the target frequency which reflects the probability of amino acid *i* ($i = 1,2,\ldots,20$) occurring at the sequence position *j* ($j = 1,2,\cdots,L$) in protein *p* during evolutionary processes. For each column in ***M*** the elements add up to one.

The target frequency is calculated from the multiple sequence alignments generated by running PSI-BLAST [35] against the NCBI's NR dataset with parameters (-j 10, -e 0.001). The target frequency of amino acid *i* in sequence position *j* is calculated as:

$$m_{i,j} = \frac{(\alpha f_{ij} + \beta g_{ij})}{(\alpha + \beta)} \tag{2}$$

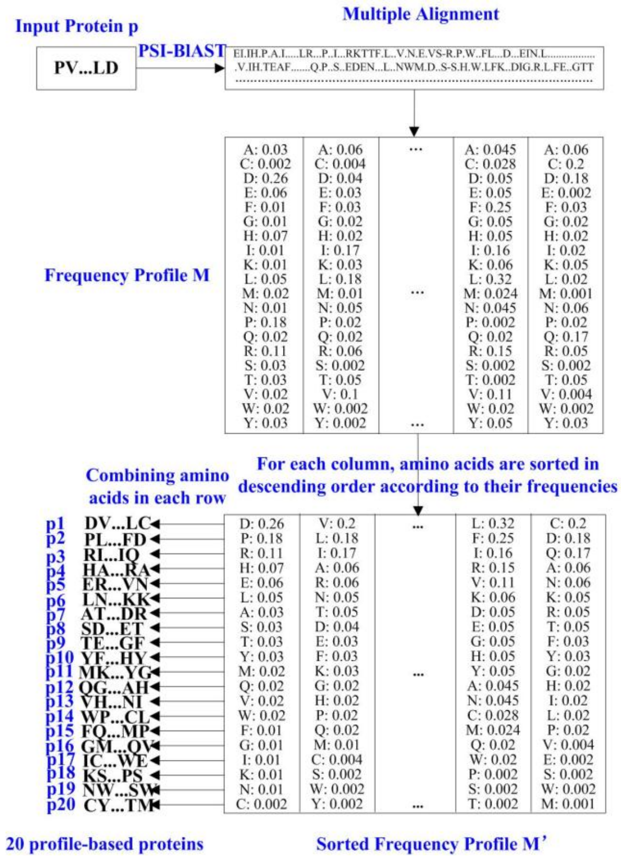
where f_{ij} is the observed frequency of amino acid *i* in column *j*; β is a free parameter set to a constant value of 10, which is initially used by PSI-BLAST, and α is the number of different amino acids in column *j* minus one. g_{ij} is the pseudo-count for amino acid *i* in protein sequence position *j*, which can be calculated as:

$$g_{ij} = \sum_{k=1}^{20} \frac{f_{kj} q_{ik}}{p_k} \tag{3}$$

where p_k is the background frequency of amino acid *k*, q_{ik} is the score of amino acid *i* being aligned to amino acid *k* in BLOSUM62 substitution matrix, which is the default score matrix of PSI-BLAST [35].

C. Profile-Based Protein Representation

Although methods using amino acid composition achieve certain degree of success, only using sequence information cannot accurately detect protein remote homology. Recent studies demonstrate the profile-based methods show better performance as the profile is a richer encoding of protein sequence than the individual sequence. However, a profile is a matrix, while a protein sequence is a string of amino acids. Therefore, the sequence-based methods cannot directly incorporate the evolutionary information in the profiles into the prediction. We propose a approach to convert the frequency profiles into a series of profile-based proteins, and then the existing sequence-based methods can be directly performed on these proteins for the prediction. The target frequencies in the frequency profiles reflect the probabilities of the corresponding amino acids appearing in the specific sequence positions. The higher the frequency is, the more likely the corresponding amino acid occurs. It is reasonable to use the *n*-th most frequent amino acids in the frequency profiles to represent the protein sequences. The following details how to convert frequency profiles into profile-based proteins.


Input Protein p
Multiple Alignment
PSI-BlAST
PV...LD
EI.IH.P.A.I.....LR...P..I...RKTTF.L..V.N.E.VS-R.P.W..FL...D...EIN.L................
.V.IH.TEAF.......Q.P..S..EDEN...L..NWM.D..S-S.H.W.LFK..DIG.R.L.FE..GTT
Frequency Profile M
For each column, amino acids are sorted in descending order according to their frequencies
Combining amino acids in each row
20 profile-based proteins
Sorted Frequency Profile M'

A: 0.03	A: 0.06	…	A: 0.045	A: 0.06
C: 0.002	C: 0.004		C: 0.028	C: 0.2
D: 0.26	D: 0.04		D: 0.05	D: 0.18
E: 0.06	E: 0.03		E: 0.05	E: 0.002
F: 0.01	F: 0.03		F: 0.25	F: 0.03
G: 0.01	G: 0.02		G: 0.05	G: 0.02
H: 0.07	H: 0.02		H: 0.05	H: 0.02
I: 0.01	I: 0.17		I: 0.16	I: 0.02
K: 0.01	K: 0.03		K: 0.06	K: 0.05
L: 0.05	L: 0.18		L: 0.32	L: 0.02
M: 0.02	M: 0.01	…	M: 0.024	M: 0.001
N: 0.01	N: 0.05		N: 0.045	N: 0.06
P: 0.18	P: 0.02		P: 0.002	P: 0.02
Q: 0.02	Q: 0.02		Q: 0.02	Q: 0.17
R: 0.11	R: 0.06		R: 0.15	R: 0.05
S: 0.03	S: 0.002		S: 0.002	S: 0.002
T: 0.03	T: 0.05		T: 0.002	T: 0.05
V: 0.02	V: 0.1		V: 0.11	V: 0.004
W: 0.02	W: 0.002		W: 0.02	W: 0.002
Y: 0.03	Y: 0.002	…	Y: 0.05	Y: 0.03

p1	DV...LC	D: 0.26	V: 0.2	…	L: 0.32	C: 0.2
p2	PL....FD	P: 0.18	L: 0.18		F: 0.25	D: 0.18
p3	RI...IQ	R: 0.11	I: 0.17		I: 0.16	Q: 0.17
p4	HA...RA	H: 0.07	A: 0.06		R: 0.15	A: 0.06
p5	ER...VN	E: 0.06	R: 0.06		V: 0.11	N: 0.06
p6	LN...KK	L: 0.05	N: 0.05		K: 0.06	K: 0.05
p7	AT...DR	A: 0.03	T: 0.05		D: 0.05	R: 0.05
p8	SD...ET	S: 0.03	D: 0.04		E: 0.05	T: 0.05
p9	TE...GF	T: 0.03	E: 0.03		G: 0.05	F: 0.03
p10	YF...HY	Y: 0.03	F: 0.03		H: 0.05	Y: 0.03
p11	MK...YG	M: 0.02	K: 0.03	…	Y: 0.05	G: 0.02
p12	QG...AH	Q: 0.02	G: 0.02		A: 0.045	H: 0.02
p13	VH...NI	V: 0.02	H: 0.02		N: 0.045	I: 0.02
p14	WP...CL	W: 0.02	P: 0.02		C: 0.028	L: 0.02
p15	FQ...MP	F: 0.01	Q: 0.02		M: 0.024	P: 0.02
p16	GM...QV	G: 0.01	M: 0.01		Q: 0.02	V: 0.004
p17	IC...WE	I: 0.01	C: 0.004		W: 0.02	E: 0.002
p18	KS...PS	K: 0.01	S: 0.002		P: 0.002	S: 0.002
p19	NW...SW	N: 0.01	W: 0.002		S: 0.002	W: 0.002
p20	CY...TM	C: 0.002	Y: 0.002	…	T: 0.002	M: 0.001

Figure 1. Flowchart of generating profile-based proteins. The multiple sequence alignment is obtained by PSI-BLAST. The frequency profile is calculated from the multiple sequence alignment. The frequencies of the 20 standard amino acids in frequency profile M are sorted in descending order and then the sorted frequency profile M' is converted into 20 profile-based proteins by combining the amino acids in each row.

Given the frequency profile ***M*** of protein *p* (equation 1), For each column in ***M***, the amino acids are sorted in descending order. Therefore, ***M*** is converted into the sorted frequency profile ***M***', and then for each row in ***M***', the amino acids are combined to produce the profile-based protein. By following this approach, the frequency profile ***M*** is converted into 20 profile-based proteins p_1, p_2, …, p_{20} (Fig. 1), which contain the evolutionary information in the frequency profile. These 20 proteins have different importance. During evolutionary process, protein *p* is preferred to transform into p1, but not preferred to transform into *p*20.

D. Local Alignemnt Kernel

The Local Alignment kernel is calculated by summing up scores obtained from the local alignments with gaps between the two sequences, computed by Smith-Waterman dynamic programming algorithm [10]. 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 implement one of the methods, namely, the LA-ekm kernel. The parameters of LA-ekm kernel take the optimal values (β = 0.5, *d* = -11, *e* = -4).

E. Support Vector Machine

Support vector machine (SVM) is a class of supervised learning algorithms first introduced by Vapnik [7]. Given a set of labelled training vectors (positive and negative input samples), SVM can learn a linear decision boundary to discriminate the two classes. The result is a linear classification rule that can be used to classify new test samples. When the samples are linearly non-separable, the kernel function can be used to map the samples to a high-order feature space in which the optimal decision boundary can be found. SVM has exhibited excellent performance in practice and has a strong theoretical foundation of statistical learning. In this study, the publicly available Gist SVM package (http://www.chibi.ubc.ca/gist/) is employed.

F. Evaluation Methodology

Because the test sets have many more negative than positive samples, simply measuring error-rates will not give a good evaluation of performance. For the cases in which the positive and negative samples are not evenly distributed, the best way to evaluate the trade-off between the specificity and sensitivity is to use a receiver operating characteristics (ROC) score [38]. A ROC score is the normalized area under a curve that plots true positives against false positives for different classification thresholds. A score of 1 denotes perfect separation of positive samples from negative ones, whereas a score of 0 indicates that none of the sequences selected by the algorithm is positive. Another performance measure is ROC50 score, which is the area under the ROC curve up to the first 50 false positives.

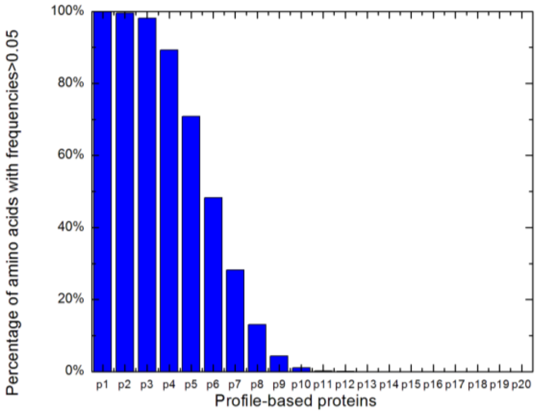


Figure 2. Illustration to show the feature of frequency profile. Percentage of amino acids with frequencies higher than 0.05 in the 20 profile-based proteins derived from SCOP 1.53 benchmark.

III. Results and Discussion

A. Profile-Based Protein Representation Can Improve the Performance of Methods based on Sequence Composition

The frequency profile of a protein *p* can be converted into 20 profile-based proteins (*p*1, *p*2, …, *p*20) by using the proposed approach (see method section for details). These 20 proteins have different importance. *p*1 is the most important protein as it is the combination of the top frequent amino acids in frequency profile, while *p*20 is the profile-based protein which protein *p* is not likely to convert into as it is the combination of the amino acids with lowest frequencies in frequency profile. If all the 20 profile-based proteins are used in the prediction, the computational cost is relatively high. In this study, only the top *n* most important profile-based proteins (*p*1,…*pn*) are used in the prediction. In order to select the value of *n*, the following experiment is conducted. The frequencies of 20 standard amino acids in each column of a frequency profiles add up to one. Therefore, the average frequency is 0.05 (1/20=0.05). If a amino acid with frequency higher than 0.05, it is likely to occur during evolutionary process, otherwise, it is not likely to occur. The percentage of the amino acids with frequencies higher than 0.05 in each profile-based protein on the SCOP 1.53 benchmark is calculated and the results are shown in Fig. 2. As shown in this figure, such amino acids are abundant in profile-based proteins *p*1, *p*2 and *p*3 (99.99%, 99.60% and 98.13%), but for the other 17 profile-based proteins, the percentage decreases significantly (from 89.28% to 0%). Therefore, in this study only the top three profile-based proteins are used in the prediction. These profile-based proteins are combined with SVM-LA [10], and the results are shown in Table I. The method performed on the top important protein *p*1 can achieve the best performance. Compared with the method performed on the raw protein sequence *p*, the performance of the proposed method can be improved by 3.7% and 13.6% in terms of average ROC score and average ROC50 score respectively, indicating that the proposed profile-based protein representation is useful for protein remote homology detection. The performance of the method performed on *p*2 is similar as the that of the method performed on the raw proteins *p*. The predictive results of the method performed on *p*3 is the lowest. These results are consistent with the different importance of the three profile-based proteins *p*1, *p*2, and *p*3.

B. Comparison with Closely Related Methods

Beside the proposed methods, several other methods attempt to predict protein remote homologies based on frequency profile. Both SVM-Top-n-gram-combine-LSA [20] and SVM-PDT-Profile [33] take the evolutionary information extracted from the into consideration. SVM-Top-n-gram-combine-LSA [20] extracts the building blocks of proteins from the frequency profiles, which can be treated as the “words” of protein language. The Latent Semantic Analysis (LSA) [8] is applied to further improve the performance of this method. SVM-PDT-Profile [33] combines the amino acid physicochemical properties in the Amino Acid Index (AAIndex) [39] with the frequency profiles for the prediction. The results of these two methods are listed in Table I. The proposed methods outperform both of the two methods, indicating that the proposed profile-based protein representation is a suitable approach to extract the evolutionary information from frequency profiles for protein remote homology detection.

TABLE I. Average ROC and ROC50 Scores over 54 Families for Different Methods

Methods	Mean ROC	Mean ROC50
SVM-LA (*p*)	0.921	0.752
SVM-LA (*p*1)	**0.958**	**0.888**
SVM-LA (*p*2)	0.898	0.770
SVM-LA (*p*3)	0.873	0.656
SVM-pairwise (*p*)	0.908	0.787
SVM-Ngram (*p*)	0.812	0.589
SVM-Top-n-gram-combine-LSA	0.939	0.767
SVM-PDT-Profile (*ß*=8, *n*=2)	0.950	0.740

p, *p*1, *p*2and *p*3 refer to the methods based on feature space derived from the raw proteins *p*, profile-based proteins *p*1, profile-based proteins *p*2 and profile-based proteins *p*3, respectively.

IV. Conclusions

Discriminative methods based on support vector machine (SVM) are the most effective and accurate methods for protein remote homology detection. The performance of the SVM-based methods depends on the kernel function, which measures the similarity between any pair of samples. Variety of kernels based on sequence composition have been proposed. However, these methods often fail to accurately predict the proteins sharing low sequence similarity. Recently, methods using the evolutionary information extracted from profiles achieve great success, such as Profile [21], SW-PSSM [22], SVM-Top-N-gram [20], SVM-ACC [23]. A key step to improve the performance of these methods is to find a suitable approach to incorporate the evolutionary information extracted from profiles into the prediction. In this article, we propose a method that can convert the frequency profile into a series of profile-based proteins. The Local Alignemnt kernel (SVM-LA [10]) is selected to demonstrate if the proposed profile-based protein representation can improve the performance of this method. Experiments on a well-known benchmark show that the methods based on the profile-based protein *p*1 and *p*2 achieve the best performance, which outperforms the original three string kernels by 3.7% and 13.6% in terms of average ROC and ROC50 scores respectively. These results are consistent with our previous findings that the top two most frequent amino acids show stronger discriminative power than the other low frequent amino acids in the frequency profiles [20]. The experimental results confirm that the proposed profile-based protein representation is a suitable approach to extract the evolutionary information from frequency profiles for protein remote homology detection.

The proposed profile-based protein representation provide a general framework to incorporate the evolutionary information in the frequency profiles into the prediction. This approach can be easily combined with sequence-based methods. With the development of the sequence-based kernels, the proposed method can be further improved. Further studies will focus on combining the profile-based protein representation with other sequence-based kernels.

ACKNOWLEDGEMENTS

This work was supported by Scientific Research Innovation Foundation in Harbin Institute of Technology (Project No. HIT.NSRIF.2013103), Natural Science Foundation of Guangdong Province (No. S2012040007390), Shenzhen International Cooperation Research Funding (No.GJHZ20120613110641217), Postdoctoral Funding (No. 2011M500669), and Shanghai Key Laboratory of Intelligent Information Processing, China (Grant No. IIPL-2012-002).

REFERENCES

[1] H. Berman, *et al.*, "The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data," *Nucleic Acids. Res.*, vol. 35, pp. D301–D303, 2007.

[2] C. H. Wu, *et al.*, "The universal protein resource (UniProt): An expanding universe of protein information," *Nucleic Acids. Res.*, vol. 34, pp. D187-D191, 2006.

[3] S. F. Altschul, *et al.*, "Basic local alignment search tool," *J Mol Biol*, vol. 215, pp. 403-410, 1990.

[4] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *J Mol Biol*, vol. 147, pp. 195-197, 1981.

[5] K. Karplus, *et al.*, "Hidden markov models for detecting remote protein homologies," *Bioinformatics*, vol. 14, pp. 846-856, 1998.

[6] B. Qian and R. A. Goldstein, "Performance of an iterated T-Hmm for homology detection," *Bioinformatics*, vol. 20, pp. 2175-2180, 2004.

[7] V. N. Vapnik, *Statistical Learning Theory*, New York, 1998.

[8] Q. W. Dong, *et al.*, "Application of latent semantic analysis to protein remote homology detection," *Bioinformatics*, vol. 22, pp. 285-290, 2006.

[9] L. Liao and W. S. Noble, "Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships," *J. Comput Biol.*, vol. 10, pp. 857-868, 2003.

[10] H. Saigo, *et al.*, "Protein homology detection using string alignment kernels," *Bioinformatics*, vol. 20, pp. 1682-1689, 2004.

[11] C. G. Nevill-Manning, *et al.*, "Highly specific protein sequence motifs for genome analysis," *Proc Natl. Acad. Sci USA* vol. 95, pp. 5865-5871, 1998.

[12] A. Ben-Hur and D. Brutlag, "Remote homology detection: A motif based approach," *Bioinformatics*, vol. 19(Suppl 1), pp. i26-i33, 2003.

[13] T. Håndstad, *et al.*, "Motif kernel generated by genetic programming improves remote homology and fold detection," *BMC Bioinformatics*, vol. 8, p. 23, 2007.

[14] C. S. Leslie, *et al.*, "Mismatch string kernels for discriminative protein classification," *Bioinformatics*, vol. 20, pp. 467-476, 2004.

[15] Y. Hou, *et al.*, "Efficient remote homology detection using local structure," *Bioinformatics*, vol. 19, pp. 2294-2301, 2003.

[16] H. Ogul and E. U. Mumcuoglu, "A discriminative method for remote homology detection based on n-peptide compositions with reduced amino acid alphabets," *BioSystems*, vol. 87, pp. 75-81, 2007.

[17] C. Leslie, *et al.*, "The spectrum kernel: A string kernel for svm protein classification," *Proc Pacific Symposium on Biocomputing*, pp. 566-575, 2002.

[18] Q. Dong, *et al.*, "A pattern-based svm for protein remote homology detection," presented at the 4th International Conference on Machine Learning And Cybernetics, GuangZhou, China, 2005.

[19] B. J. Webb-Robertson, *et al.*, "SVM-BALSA: Remote homology detection based on Bayesian sequence alignment," *Computational Biology and Chemistry*, vol. 29, pp. 440-443, 2005.

[20] B. Liu, *et al.*, "A discriminative method for protein remote homology detection and fold recognition combining Top-n-grams and latent semantic analysis," *BMC Bioinformatics*, vol. 9, p. 510, 2008.

[21] R. Kuang, *et al.*, "Profile-based string kernels for remote homology detection and motif extraction " *J. Bioinform. Comput. Biol.*, vol. 3, pp. 527-550, 2005.

[22] H. Rangwala and G. Karypis, "Profile-based direct kernels for remote homology detection and fold recognition," *Bioinformatics*, vol. 21, pp. 4239-4247, 2005.

[23] X. Liu, *et al.*, "Protein remote homology detection based on auto-cross covariance transformation," *Computers in Biology and Medicine*, vol. 41, pp. 640-647, 2011.

[24] J. Söding, "Protein homology detection by HMM–HMM comparison," *Bioinformatics*, vol. 21, pp. 951-960, 2005.

[25] R. I. Sadreyev, *et al.*, "COMPASS server for homology detection: improved statistical accuracy, speed and functionality," *Nucleic Acids Research*, vol. 37, pp. W90-W94, 2009.

[26] D. Gront, *et al.*, "BioShell Threader: protein homology detection based on sequence profiles and secondary structure profiles," *Nucleic Acids Research*, vol. 40, pp. W257-W262, 2012.

[27] K. Tomii and Y. Akiyama, "FORTE: A profile-profile comparison tool for protein fold recognition.," *Bioinformatics*, vol. 20, pp. 594-595, 2004.

[28] M. Margelevicius and M. L. C. Venclovas, "COMA server for protein distant homology search," *Bioinformatics*, vol. 26, pp. 1905-1906, profile-profile alignment 2010.

[29] L. A. Kelley and M. J. Sternberg, "Protein structure prediction on the Web: A case study using the Phyre server," *Nat. Protoc.*, vol. 4, pp. 363-371, 2009.

[30] A. Lobley, *et al.*, "pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination," *Bioinformatics*, vol. 25, pp. 1761-1767, 2009.

[31] B. W. Brandt and J. Heringa, "webPRC: The Profile Comparer for alignment-based searching of public domain databases," *Nucleic Acids Res.*, vol. 37, pp. W48-W52, 2009.

[32] T. Damoulas and M. A. Girolami, "Probabilistic multi-class multi-kernel learning: on protein fold recognition and remote homology detection," *Bioinformatics*, vol. 24, pp. 1264-1270, 2008.

[33] B. Liu, *et al.*, "Using amino acid physicochemical distance transformation for fast protein remote homology detection," *PLoS ONE*, vol. 7, p. e46633, 2012.

[34] I. Melvin, *et al.*, "Detecting remote evolutionary relationships among proteins by large-scale semantic embedding," *PLoS Computational Biology*, vol. 7, p. e1001047, 2011.

[35] S. F. Altschul, *et al.*, "Gapped blast and psi-blast: A new generation of protein database search programs," *Nucleic Acids Res*, vol. 25, pp. 3389-3402, 1997.

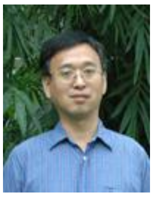
[36] T. Lingner and P. Meinicke, "Remote homology detection based on oligomer distances," *Bioinformatics*, vol. 22, pp. 2224-2231, 2006.

[37] S. E. Brenner, *et al.*, "The ASTRAL compendium for sequence and structure analysis," *Nucleic Acids Res*, vol. 28, pp. 254-256, 2000.

[38] M. Gribskov and N. L. Robinson, "Use of receiver operating characteristic (Roc) analysis to evaluate sequence matching," *Comput Chem*, vol. 20, pp. 25-33, 1996.

[39] S. Kawashima, *et al.*, "AAindex: Amino acid index database, progress report 2008," *Nucleic Acids Res.*, vol. 36, pp. D202-D205, 2008.

Bin Liu He is an assistant professor at Harbin Institute of Technology Shenzhen Graduate School. He received his Ph.D from Harbin Institute of Technology in 2010, and then worked at The Ohio State University as a post doctoral researcher during 2011. His research interests are Bioinformatics, natural language processing, data mining.



Xiaolong Wang He is a professor at Harbin Institute of Technology Shenzhen Graduate School. His research interests are Bioinformatics, natural language processing, data mining.

Ruifeng Xu He is an associate professor at Harbin Institute of Technology Shenzhen Graduate School. His research interests are Bioinformatics, natural language processing, data mining.

Buzhou Tang He is a post doctoral researcher at Harbin Institute of Technology Shenzhen Graduate School. His research interests are Bioinformatics, natural language processing, data mining.