Engineering, **2013, 5, 149-153**
http://dx.doi.org/10.4236/eng.2013.510B032 Published Online October 2013 (http://www.scirp.org/journal/eng)

Using Chou's Pseudo Amino Acid Composition for Protein Remote Homology Detection

Bin Liu[1,2,3], Xiaolong Wang[1,2]
[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
Email: bliu@insun.hit.edu.cn, wangxl@insun.hit.edu.cn

Received April 2013

ABSTRACT

Protein remote homology detection is a key problem in bioinformatics. Currently, the discriminative methods, such as Support Vector Machine (SVM), can achieve the best performance. The most efficient approach to improve the performance of the SVM-based methods is to find a general protein representation method that is able to convert proteins with different lengths into fixed length vectors and captures the different properties of the proteins for the discrimination. The bottleneck of designing the protein representation method is that native proteins have different lengths. Motivated by the success of the pseudo amino acid composition (PseAAC) proposed by Chou, we applied this approach for protein remote homology detection. Some new indices derived from the amino acid index (AAIndex) database are incorporated into the PseAAC to improve the generalization ability of this method. Our experiments on a well-known benchmark show this method achieves superior or comparable performance with current state-of-the-art methods.

Keywords: Protein Remote Homology; Support Vector Machine; Pseudo Amino Acid Composition; Protein Representation

1. Introduction

Protein remote homology detection, referring to the detection of evolutional homology in proteins with low similarities, is a challenging problem in bioinformatics, which has been intensively studied for a decade. Many computational methods have been proposed to address this problem, which can be split into three groups: pairwise comparison methods, generative models and discriminative algorithms. Pairwise comparison methods measure the pairwise similarities between protein sequences, such as pairwise method [1] and Smith-Waterman dynamic programming algorithm [2]. Generative models 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 [3]. Recent methods have applied the discriminative algorithms for accurate remote homology detection. Different from the generative methods, the discriminative methods lean a combination of the features that can discriminate the protein families. Among these methods, the top-performing methods use the support vector machines (SVM) [4] to build the discriminative framework. The core component in the SVM is the calculation of the kernel functions, which measure the difference between any two pair of samples. For example, LA kernel [5] measures the similarity between a pair of proteins by taking all the optimal local alignment scores with gaps between all possible subsequences into account. SVM-PDT [6] takes the sequence order information of the proteins into account by combining the amino acid physicochemical distance transformation and different amino acid indices derived from the AAIndex database [7]. Some top-performing methods employ the evolutional information extracted from the profiles. These methods need an additional alignment step to generate the profiles by searching against a non-redundant database, which leads to higher computational cost. For example, Top-*n*-grams extract the profile-based patterns by considering the most frequent elements in the profiles [8].

A key step to improve the performance of the SVM-based methods is to find a fast and accurate representation of protein sequence. Previous studies show that the sequence order effects are relevant for remote homology detection [9]. The difficulty to include the sequence order information into the prediction is that protein sequence lengths vary widely. The pseudo amino acid composition

(PseAAC) was proposed by Chou [10]. Motivated by the success of PseAAC, we applied this approach for protein remote homology detection. Some new indices derived from the AAIndex database are incorporated into the PseAAC to improve the generalization ability of this method.

2. Methods

2.1. Benchmark Dataset

A common benchmark [1] 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 [5,9,11], 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 [12] 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.

2.2. Amino Acid Indices

The Amino Acid Index (AAIndex) [7] is a database of numerical indices representing various physicochemical and biochemical properties of amino acids and pairs of amino acids (http://www.genome.jp/aaindex/). There are three sections in the latest version of the database (version 9): AAIndex1, AAIndex2 and AAIndex3. AAIndex1 contains 544 indices; AAIndex2 has 94 amino acid mutation matrices; AAIndex3 has 47 statistical protein contact potential matrices. Because AAIndex2 and AAIndex3 are matrices, they are not suitable for PseAAC. Therefore, the AAIndex1 is selected for protein transformation step. After removing the incomplete data and the indices with all zeros in AAIndex1, 531 indices are selected for the physicochemical property distance transformation.

2.3. Combining AA Indices with Pseudo Amino Acid Composition

The pseudo amino acid composition (PseAAC) was proposed by Chou [10], which takes the sequence order information into account. PseAAC has been applied to successfully solve many important problems in computational biology, such as predicting enzymes and their family/sub-family classification [13], protein subcellular location prediction [14], predicting protein subnuclear localization [15], predicting membrane proteins and their types discrimination of outer membrane proteins [16,17], identifying proteases and their types [18], predicting protein quaternary structural attributes [19,20], fold pattern prediction [21,22], and many other tasks.

In this study, we employ the concept of PseAAC for protein remote homology detection. In the original PseAAC, it only uses three indices, including the hydrophobicity index, hydrophilicity index, and side-chain mass index. Because protein remote homology detection is a more difficult problem, proteins in the dataset only have very low sequence similarity. Only these three indices are not enough to capture the different properties of various proteins. Therefore, we extend the PseAAC by using all the meaningful 531 indices extracted from the AAIndex database, which describe the properties of the 20 standard amino acids in different aspects. The proposed method is called PseAACIndex.

The detailed process of the PseAACIndex is shown in the following.

Given a protein sequence with L amino acids:

$$A_1\ A_2\ A_3\ A_4\ A_5\ A_6.....A_L \tag{1}$$

where A_1 is the amino acid at protein chain position 1, A_2 is the amino acid at protein chain position 2 and so forth. Given an amino acid index j in AAIndex1, each protein sequence is converted into a series of numbers by using the amino acid index j.

All the 531 indices are subjected to a standard conversion by the following equation:

$$I_j(A_i) = \frac{I_j^{\wedge}(A_i) - \sum_{m=1}^{20} \frac{I_j^{\wedge}(R_m)}{20}}{\sqrt{\frac{\sum_{k=1}^{20} (I_j^{\wedge}(R_k) - \sum_{m=1}^{20} \frac{I_j^{\wedge}(R_m)}{20})^2}{20}}} \tag{2}$$

where $I^{\wedge}(A_i)$ represents the raw physicochemical property value of amino acid A_i in index j, R_m (m = 1, 2, 3, 4, …, 20) represents the 20 standard amino acids.

The sequence order information associated with index j can be approximately reflected with the order-correlated factor as defined below:

$$\partial_\lambda^j = \frac{\sum_{i=1}^{L-\lambda} (I_j(A_i) - I_j(A_{i+\lambda}))^2}{L-\lambda} \tag{3}$$

where λ is the distance between two amino acids along the protein chain.

Let us use the concept of the PseAAC to formulate the amino acid composition by using the order-correlated factor calculated by Equation (3). Given an index j, the protein sequence can be converted into a 20 + λ dimensional vector:

$$\mathbf{X}^j = [x_1, x_2, \ldots, x_{20}, \ldots, x_{20+\lambda}] \quad (4)$$

where

$$x_u = \begin{cases} \dfrac{f_u}{\sum_{m=1}^{20} f_m + \sum_{i=1}^{\lambda} \partial_i^j}, (1 \le u \le 20) \\ \dfrac{\partial_{u-20}^j}{\sum_{m=1}^{20} f_m + \sum_{i=1}^{\lambda} \partial_i^j}, (20+1 \le u \le 20+\lambda) \end{cases} \quad (5)$$

where f_m is the normalized occurrence frequency of the 20 standard amino acids in a protein, ∂_i^j is the i-tier sequence correlation factor calculated by equation 3. The first 20 elements represent the effect of the amino acid composition, and the elements from 20 + 1 to 20 + λ represent the effect of sequence order.

In this study, 531 indices are extracted from the AAIndex database. Therefore, by using the above approach, a protein can be represented as following vector:

$$\mathbf{X} = [\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^j, ..., \mathbf{x}^{531}] \quad (6)$$

where $\mathbf{X}^j$ is calculated by Equations (4) and (5). Therefore, the dimension of the final vector is 531*(20 + λ).

2.4. Support Vector Machine

Support vector machine (SVM) is a class of supervised learning algorithms first introduced by Vapnik [4]. 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. The SVM parameters are used by default of the Gist Package except that the kernel function is set as radial basis function.

2.5. Evaluation Methodology

Because the test sets have 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 [23]. 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.

3. Results and Discussion

3.1. λ Value Has Minior Impact on the Performance of PseACC-AAIndex

In our method, there is a parameter λ, which would impact on the performance of PseAACIndex (see method section for more information). λ can be any integer between 1 and L-1, where L is the shortest protein sequence in the dataset. The average ROC scores obtained by using different λ values are shown in **Figure 1**. As we can see from the figure, the λ value has little impact on the performance. PseAACIndex with different λ values show similar results. Here, the λ value of 5 is used in this study, because of this value, PseAACIndex can achieve the best performance with shorter feature vectors and lower computational cost.

3.2. Comparison with Other Sequence-Based Methods

In order to compare the proposed PseAACIndex method with other relevant protein remote homology detection methods, the PseAACIndex was evaluated on the widely used SCOP 1.53 dataset to give an unbiased comparison with prior methods that are based on sequence composition information.

Although previous study tuned both the features and SVM parameters for each protein family, in order to evaluate the robustness and generalization of the PseACC vectorization approach, no feature selection was performed to select the best features for the proteins or

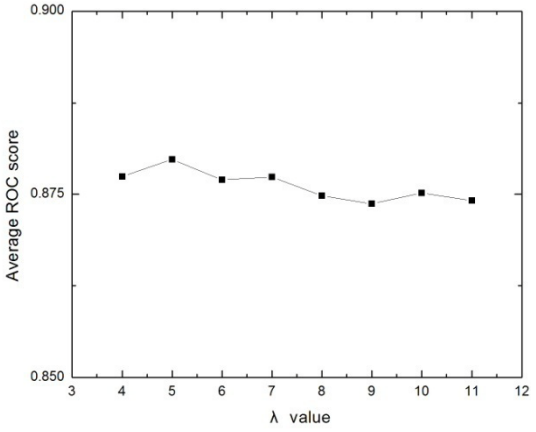

Figure 1. The impact of λ on the average ROC score obtained by PseACCIndex.

the families. All the 531 amino acid indices were used for predicting each family.

The predictive results of different sequence-based methods are listed in **Table 1**. SVM-Ngram, SVM-Pattern, and SVM-Motif are based on three different building blocks of proteins. Mismatch method allows a given number of mismatches between the substrings of the proteins. SVM-LA is based on the pairwise similarity scores. The performance of PseAACIndex is highly comparable with SVM-LA and outperforms other methods in terms of both ROC and ROC50 scores, indicating that the proposed PseAACIndex approach is an efficient method for protein remote homology detection.

4. Conclusion

In this study, inspired by the success of PseAAC, we combined the PseAAC with various amino acid indices extracted from the AAIndex database for protein remote homology detection. It took both the sequence-order information and the amino acid physicochemical properties extracted from the AAIndex database into consideration. Experimental results demonstrated that this approach was useful for protein remote homology detection and showed better predictive results than the compared methods.

Table 1. Results of different methods for protein remote homology detection.

Average ROC and ROC50 scores			
Methods	**ROC**	**ROC50**	**Source**
PseAACIndex (λ = 5)	0.880	0.620	This study
SVM-Ngram	0.791	0.584	[24]
SVM-Pattern	0.835	0.589	[24]
SVM-LA(β = 0.5)	0.925	0.649	[5]
Mismatch	0.872	0.400	[25]
SVM-Motif	0.814	0.616	[24]

5. Acknowledgements

We would like to thank Professor Kuo-Chen Chou at Gordon Life Science Institute for his helpful suggestions on this manuscript. This work was supported by the National Natural Science Foundation of China (No. 61173075, 61003090 and 60973076), the Project HIT. NSRIF.2013103 supported by Natural Scientific Research Innovation Foundation in Harbin Institute of Technology, Natural Science Foundation of Guangdong province (No.S2012040007390), and Shanghai Key Laboratory of Intelligent Information Processing, China (Grant No.IIPL-2012-002).

REFERENCES

[1] L. Liao and W. S. Noble, "Combining Pairwise Sequence Similarity and Support Vector Machines for Detecting Remote Protein Evolutionary and Structural Relationships," *Journal of Computational Biology*, Vol. 10, No. 6, 2003, pp. 857-868. http://dx.doi.org/10.1089/106652703322756113

[2] T. F. Smith and M. S. Waterman, "Identification of Common Molecular Subsequences," *Journal of Molecular Biology*, Vol. 147, No. 1, 1981, pp. 195-197. http://dx.doi.org/10.1016/0022-2836(81)90087-5

[3] B. Qian and R. A. Goldstein, "Performance of an Iterated T-Hmm for Homology Detection," *Bioinformatics*, Vol. 20, No. 14, 2004, pp. 2175-2180. http://dx.doi.org/10.1093/bioinformatics/bth181

[4] V. N. Vapnik, "Statistical Learning Theory," 1998.

[5] H. Saigo, *et al*., "Protein Homology Detection Using String Alignment Kernels," *Bioinformatics*, Vol. 20, No. 11, 2004, pp. 1682-1689. http://dx.doi.org/10.1093/bioinformatics/bth141

[6] B. Liu, *et al*., "Using Amino Acid Physicochemical Distance Transformation for Fast Protein Remote Homology Detection," *PLoS ONE*, Vol. 7, No. 9, 2012, p. e46633. http://dx.doi.org/10.1371/journal.pone.0046633

[7] S. Kawashima, *et al*., "AAindex: Amino Acid Index Database, Progress Report 2008," *Nucleic Acids Research*, Vol. 36, No. Database, 2008, pp. D202-D205.

[8] 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, 2008, p. 510. http://dx.doi.org/10.1186/1471-2105-9-510

[9] T. Lingner and P. Meinicke, "Remote Homology Detection Based on Oligomer Distances," *Bioinformatics*, Vol. 22, No. 18, 2006, pp. 2224-2231. http://dx.doi.org/10.1093/bioinformatics/btl376

[10] K. C. Chou, "Prediction of Protein Cellular Attributes Using Pseudo Amino Acid Composition," *Proteins*: *Structure*, *Function*, *and Bioinformatics*, Vol. 43, 2001, pp. 246-255. http://dx.doi.org/10.1002/prot.1035

[11] Q. W. Dong, *et al*., "Application of Latent Semantic Analysis to Protein Remote Homology Detection," *Bioinformatics*, Vol. 22, No. 3, 2006, pp. 285-290. http://dx.doi.org/10.1093/bioinformatics/bti801

[12] S. E. Brenner, *et al*., "The ASTRAL Compendium for Sequence and Structure Analysis," *Nucleic Acids Research*, Vol. 28, No. 1, 2000, pp. 254-256. http://dx.doi.org/10.1093/nar/28.1.254

[13] Y. D. Cai and K. C. Chou, "Predicting Enzyme Subclass by Functional Domain Composition and Pseudo Amino Acid Composition," *Journal of Proteome Research*, Vol. 4, 2005, pp. 967-971. http://dx.doi.org/10.1021/pr0500399

[14] Y. D. Cai and K. C. Chou, "Nearest Neighbour Algorithm for Predicting Protein Subcellular Location by Combining Functional Domain Composition and Pseudoamino Acid Composition," *Biochemical and Biophysical Research Communications*, Vol. 305, 2003, pp. 407-411. http://dx.doi.org/10.1016/S0006-291X(03)00775-7

[15] H. B. Shen and K. C. Chou, "Predicting Protein Subnuc-

lear Location with Optimized Evidence-Theoretic K-Nearest Classifier and Pseudo Amino Acid Composition," *Biochemical and Biophysical Research Communications*, Vol. 337, 2005, pp. 752-756. http://dx.doi.org/10.1016/j.bbrc.2005.09.117

[16] Y. D. Cai and K. C. Chou, "Predicting Membrane Protein Type by Functional Domain Composition and Pseudo Amino Acid Composition," *Journal of Theoretical Biology*, Vol. 238, 2006, pp. 395-400. http://dx.doi.org/10.1016/j.jtbi.2005.05.035

[17] K. C. Chou and H. B. Shen, "MemType-2L: AWEB Server for Predicting Membrane Proteins and Their Types by Incorporating Evolution Information through Pse-PSSM," *Biochemical and Biophysical Research Communications*, Vol. 360, 2007, pp. 339-345. http://dx.doi.org/10.1016/j.bbrc.2007.06.027

[18] K. C. Chou and H. B. Shen, "ProtIdent: A Web Server for Identifying Proteases and Their Types by Fusing Functional Domain and Sequential Evolution Information," *Biochemical and Biophysical Research Communications*, Vol. 376, 2008, pp. 321-325. http://dx.doi.org/10.1016/j.bbrc.2008.08.125

[19] K. C. Chou and Y. D. Cai, "Predicting Protein Quaternary Structure by Pseudo Amino Acid Composition," *Proteins*: *Structure*, *Function*, *and Bioinformatics*, Vol. 53, 2003, pp. 282-289. http://dx.doi.org/10.1002/prot.10500

[20] H. B. Shen and K. C. Chou, "QuatIdent: A Web Server for Identifying Protein Quaternary Structural Attribute by Fusing Functional Domain and Sequential Evolution Information," *Journal of Proteome Research*, Vol. 8, 2009, pp. 1577-1584. http://dx.doi.org/10.1021/pr800957q

[21] H. B. Shen and K. C. Chou, "Ensemble Classifier for Protein Fold Pattern Recognition," *Bioinformatics*, Vol. 22, 2006, pp. 1717-1722. http://dx.doi.org/10.1093/bioinformatics/btl170

[22] H. B. Shen and K. C. Chou, "Predicting Protein Fold Pattern with Functional Domain and Sequential Evolution Information," *Journal of Theoretical Biology*, Vol. 256, 2009, pp. 441-446. http://dx.doi.org/10.1016/j.jtbi.2008.10.007

[23] M. Gribskov and N. L. Robinson, "Use of Receiver Operating Characteristic (Roc) Analysis to Evaluate Sequence Matching," *Computational Chemistry*, Vol. 20, No. 1, 1996, pp. 25-33. http://dx.doi.org/10.1016/S0097-8485(96)80004-0

[24] Q. Dong, *et al*., "Protein Remote Homology Detection Based on Binary Profiles," *Proceedings of* 1*st International Conference on Bioinformatics Research and Development* (*BIRD*), Germany, 2007, pp. 212-223.

[25] C. S. Leslie, *et al*., "Mismatch String Kernels for Discriminative Protein Classification," *Bioinformatics*, Vol. 20, No. 4, 2004, pp. 467-476. http://dx.doi.org/10.1093/bioinformatics/btg431

