MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **67** (2012) 833-843

ISSN 0340 - 6253

A Novel Method for Protein Function Prediction Based on Sequence Numerical Features

Ang Yang[1*], Renfa Li[1*], Wen Zhu[1], Guangxue Yue[2]

[1] School of Information Science and Engineering, Hunan University,
Changsha Hunan, 410082, China

[2] College of Mathematics and Information Engineering, Jiaxing University,
Jiaxing Zhejiang, 314001, China

(Received May 23, 2011)

Abstract

Compared with costly and time-consuming biological experiments, computational approaches to predict protein functions are easier and more cost-efficient. In this work, a feature vector constructed by extracting numerical features from sequences based on hydrophobicity, polarity and charge properties, and a function possibility of sequence are proposed. Then the feature vector and function possibility are used to predict protein function with k-nearest neighbors algorithm (KNN). Our method avoids some problems of sequence similarity based methods, because it has involved both local and global information of sequences. The results of our experiments show that our method is more efficient.

1 Introduction

An essential goal of bioinformatics is to predict the functions of unknown proteins. Since it is expensive and time-consuming to determine the functions of proteins through experiments, it is therefore important and essential to study computational approaches.

Currently, many methods have been developed to predict the functions of proteins. Some methods are based on sequence similarity [1], for example, by using BLAST[2], FASTA[3], researchers can carry out a sequence similarity search to find similar proteins or annotation information in public databases [4]. Some methods are based on structure similarity.

[*] Corresponding author. Fax: +86 731 88821417
E-mail address: jt_angyang @163.com (A.Yang), jt_lrf @163.com (R F. Li)

Kawabata and Eidhammer used structure similarity to predict functions. The protein-protein interaction approach was used in the prediction of protein functions [5,6]. Vazquez proposed assigning proteins functional class based on the network of physical interactions, which are determined by minimizing the number of protein interactions among different functional categories [7]. Function assignment is proteome-wide，determined by the global connectivity pattern of the protein network. The approach results in multiple functional assignments, which is an equivalent solution. The method of combining of sequence and structural features was proposed. Pugalenthi presented a SVM method for the identification of catalytic residues using sequence and structural features [8]. In addition, many machine-learning algorithms have been used in function prediction, such as support vector machines (SVM) [9],neural networks [10-12], Naive Bayes classifiers [13,14] and so on. Among them SVM is most widely used. But for all of the machine-learning algorithms, the results differ from the training sets.

The features extracted from the sequence play an important role in function prediction. Many researchers have acquired a variety of features extracted from sequences which can be used in function prediction. Jong Kyoung Kim presented a feature extraction method from protein sequence, which employs local and global pair-wise sequence alignment scores as well as composition-based features [15]. Five different features are used for training support vector machines (SVMs) separately and a weighted majority voting makes a final decision. The accuracy reached 88.53% when it was used in prediction of subcellular localization of proteins. In Gao’s method, a combined feature of primary sequence defined as a 430D (dimensional) vector including 20 amino acid compositions, 400 dipeptide compositions and 10 physicochemical properties, was utilized to predict the protein subcellular location [16]. Pufeng Du proposed a method to predict C-to-U RNA editing sites using only nucleotide sequence features [17]. Li and Liao proposed a global encoding method of protein sequence (GE) to describe global information of amino acid sequence, and assigned protein functional class using nearest neighbor algorithm (NNA) [18]. In Lee’s method, thirty-three features that represent subtle differences in local regions and full regions of the protein sequences were introduced. Those features were extracted from the sequences based on the transition of negatively and positively charged residues, which depends on the importance of negatively/positively charged residues [4]. Liao aimed at choosing some nearest samples according to their length for identifying protein function, irrespective of sequence and structural similarities. He proposed a method for data selection and used Nearest neighbor algorithm(NNA) to predict the protein function [19].

In this letter, new features extracted from the sequences based on the transition between different classes of amino acids are introduced, and then used to predict protein function with the algorithm of k-nearest neighbors. The results show that our method is effective.

2 Dataset

We download the 1377 protein sequences from ftp://ftpmips.gsf.de/yeast/, which were extracted from the dataset of the report of Vazquez [7]. The seventeen functional categories of all proteins are presented in Table 1 [18].

Table 1 The numbers of each functional class in dataset

Functional class	Number	Functional class	Number
Metabolism	408	Protein fate (folding, modification, destination)	452
Energy	95	Cell cycle and DNA processing	441
Development (systemic)	26	Protein with binding function of cofactor requirement	458
Cell type differentiation	204	Cellular transport, transportfacilities and transport routes	331
Protein synthesis	98	Regulation of metabolism and protein function	115
Interaction with the environment	172	Cellular communication/signal transduction mechanism	110
Cell fate	143	Cellrescue,defense and virulence	201
Biogenesis of cellular components	324	Transposable elements,viral and plasmid proteins	5
Transcription	427		

From this dataset, we randomly choose M proteins as the known-functions samples, and then choose M1 samples from the remaining proteins also at random. M1 proteins will be used as the prediction proteins in our study. At last, we use the jackknife methods to test the whole dataset.

3 Methods

3.1 Extract new features from protein sequences

Cai supported that hydrophobicity, polarity, and charge properties play greater roles than other features [20]. We classify the 20 residues into five different classes based on their physiochemical characteristics, such as hydrophobic property, polarity, acid-base properties, and so on. Also, acid-base properties are classified as negatively/positively charged residues in some papers [4].

neutral non-polar hydrophobic amino	$\underline{A}$={AVLIFPG}
neutral polar hydrophilic amino	$\underline{B}$={QSTCN}
neutral polar hydrophobic amino	$\underline{C}$={MWY}
acid-hydrophilic amino(negatively charged residues)	$\underline{D}$={DE}
Base-hydrophilic amino (positively charged residues)	$\underline{E}$={ KRH}

We extract some features from the protein sequences using the following method.

$R\underline{AA}$ was defined as

$$R\underline{AA} = \#\underline{AA}/n \qquad (1)$$

n is the total number of amino acids in a sequence, $\#\underline{AA}$ is the total number of continuous changes from $\underline{A}$ to $\underline{A}$,

$$R\underline{AB} = \#\underline{AB}/n \qquad (2)$$

$\#\underline{AB}$ is the total number of continuous changes from $\underline{A}$ to $\underline{B}$ or vice versa. Similar to $R\underline{AA}$, $R\underline{AB}$, we can get $R\underline{AC}$,$R\underline{AD}$,$R\underline{AE}$,$R\underline{BB}$,$R\underline{BC}$,$R\underline{BD}$,$R\underline{BE}$,$R\underline{CC}$,$R\underline{CD}$,$R\underline{CE}$,$R\underline{DD}$,$R\underline{DE}$, $R\underline{EE}$, a total of 15 kinds of global numerical features.

To account for local region information, we can divide every protein sequence into L parts, let $\lceil n/L \rceil$ be the length of the previous L-1 parts, and the rest of amino can be put in the L-th part. $\lceil n/L \rceil$=ceiling (n/L), the ceiling function returns the next greater integer. So for every part, we can get 15 number features and total L*15 local number features. For example

$$R\underline{BB}(i) = \#\underline{BB}(i)/n(i) \quad i = 1,2,3,\ldots,L \qquad (3)$$

$$R\underline{BC}(i) = \#\underline{BC}(i)/n(i) \quad i = 1,2,3,\ldots,L \qquad (4)$$

Coupled with global features, we can get a total of 15*L+15 features for every sequence. The dimension of the feature vector is 15*L+15.

3.2 Function Prediction Based on k-nearest Neighbor (KNN)

After get a W(=15*L+15) dimensional vector for every protein sequence, we work out similarities of different sequences by calculating the distances, such as Euclidean distance and Hamming distance and so on. In this work, Euclidean distance is used as the distance metric, as formula (5), where V_i is the vector of the i-th sequence, while V_j is the vector of the j-th sequence, V_i^s is the s-th elements of the vector V_i, and d_{ij} is the distance between V_i and V_j.

$$d_{ij} = \sqrt{\sum_{s=1}^{W} (V_i^s - V_j^s)^2} \qquad (5)$$

Assuming that there are X sample sequences whose functions have been known, we can get the X distance values between Q and the X sample sequences for a testing sequence Q. Then we select the K sequences, which are close to Q sequence.

And f_m is defined as the possibility of the Q sequence, which has the function of m, m=1,2,..., 17. There are two methods to calculate $f_{m,}$ m=1,2,..., 17.

I: Count the number of sequences which have the function of m in the K sample sequences, assign this number to f_m.

II: Order the K sequences according to the K distance values from small to large. If the 1-st sequence has the function m, f_m=f_m+K, if not, f_m =f_m; if the 2nd sequence has the function m, f_m=f_m+K-2+1, if not, f_m=f_m; if the k-th sequence has the function m, f_m=f_m+K-k+1,if not ,f_m=f_m; The rest can be done in the same manner. Then we can get the values of all the $f_{m,}$ m=1,2,...,17.

Next we can use the value of f_m to predict the functions of sequence Q. Because every sequence has more than one function, so we can also predict that Q has more than one function. We can take the one, two, three, four biggest values of f_m corresponding functions, and predict them as functions of Q sequence. This is due to the average value of all the sequences functions which is 4, and the largest is 8 in our dataset [3].

4 Results and Discussion

All of the feature vectors are calculated by our new method, and KNN is used in prediction parts of the dataset. After a lot of experiments, we find that L=20 is the best choice for the dataset. Table 2, 3 and 4 are the results of experiments when the two, three, four largest values are taken from corresponding functions that are predicted as the function of sequences respectively. Figure 1 show the results of experiments when the one, two, three, four largest values are selected for predicting the functions. Table 5 shows that the accuracy of our method is better than Vazquez [7] and Li Xi [18] in most indicators. All the data in the five tables are the average values of fifty experiments. Vazquez's global optimization method (GOM) was based on the protein-protein network, so it needs the protein-protein interaction data to use this method, and it can't handle the problem if the protein has only one interaction pattern. Our method just only used the information got from the primitive sequence, which can complete all the conditions. Li Xi's method should encode the sequences before extracting features, but our approach extracts features directly, eliminating the need for encoding process. And they used the nearest neighbor algorithm (NNA), the result of which is not as good as KNN. We test the whole dataset by jackknife method. The result in table 6 shows that our method is also effective for the whole dataset.

Table 2 While L=20,the accuracy of k=1,...,K, for different M and M1,taking the biggest two values

k	10	20	30	40	50	60	70	80	MAX	MAX-K
M=670,M1=240	0.617	0.583	0.608	0.638	0.627	0.571	0.617	0.571	0.652	17
M=247,M1=100	0.540	0.600	0.595	0.595	0.595	0.595	0.610	0.555	0.650	15
M=159,M1=64	0.586	0.555	0.586	0.602	0.594	0.578	0.539	0.594	0.695	49
M=99,M1=40	0.563	0.575	0.588	0.625	0.613	0.538	0.575	0.525	0.738	66
M=63,M1=30	0.617	0.633	0.700	0.683	0.550	0.633	0.000	0.000	0.700	30
M=34,M1=15	0.367	0.600	0.533	0.000	0.000	0.000	0.000	0.000	0.667	19

Table 3 While L=20,the accuracy of k=1,...,K, for different M and M1,taking the biggest three values

k	10	20	30	40	50	60	70	80	MAX	MAX-K
M=670,M1=240	0.702	0.754	0.754	0.752	0.750	0.729	0.765	0.725	0.773	63
M=247,M1=100	0.705	0.735	0.740	0.680	0.690	0.710	0.745	0.745	0.815	51
M=159,M1=64	0.719	0.758	0.781	0.781	0.648	0.734	0.719	0.680	0.828	79
M=99,M1=40	0.550	0.750	0.588	0.725	0.700	0.725	0.738	0.575	0.838	32
M=63,M1=30	0.700	0.617	0.767	0.650	0.683	0.683	0.000	0.000	0.833	41
M=34,M1=15	0.667	0.733	0.533	0.000	0.000	0.000	0.000	0.000	0.833	18

Table 4 While L=20,the accuracy of k=1,...,K, for different M and M1,taking the biggest four values

k	10	20	30	40	50	60	70	80	MAX	MAX-K
M=670,M1=240	0.742	0.808	0.815	0.848	0.835	0.817	0.819	0.806	0.871	44
M=247,M1=100	0.735	0.800	0.780	0.805	0.820	0.830	0.810	0.735	0.865	63
M=159,M1=64	0.758	0.813	0.852	0.789	0.805	0.836	0.773	0.820	0.875	30
M=99,M1=40	0.800	0.838	0.863	0.800	0.850	0.863	0.813	0.850	0.938	44
M=63,M1=30	0.800	0.883	0.833	0.783	0.750	0.850	0.000	0.000	0.883	54
M=34,M1=15	0.800	0.600	0.733	0.000	0.000	0.000	0.000	0.000	0.900	24

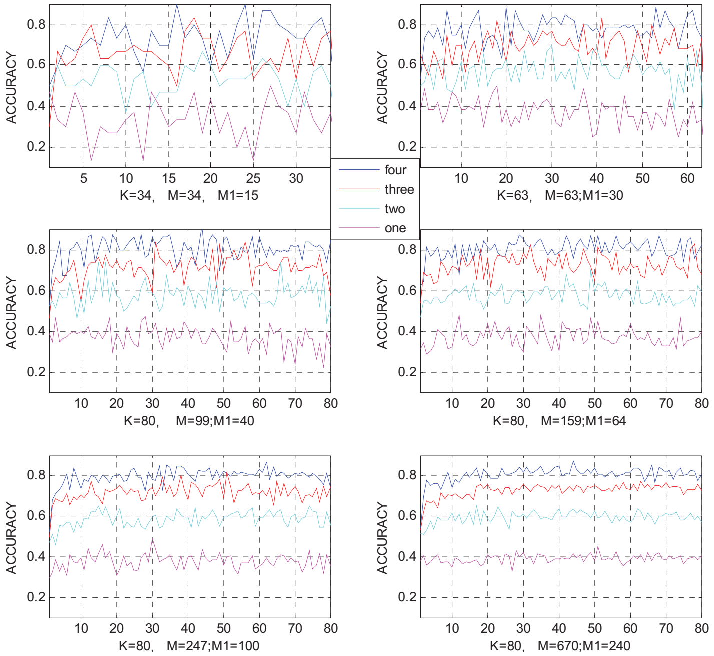

Figure 1 While L=20,different M and M1,the accuracy for k=1,2,...,K

Table 5 The accuracy comparison of our method and other methods

Accuracy	M=670	M=247	M=159	M=99	M=63	M=34
	M1=240	M1=100	M1=64	M1=40	M1=30	M1=15
Alexei.V[a]	–	0. 610	0. 760	0. 770	0. 860	0. 890
Xi Li[b]	0. 603	0. 665	0. 660	0. 780	0. 767	0. 747
Our (two)	0. 652	0. 650	0. 695	0. 738	0. 700	0. 667
Our(three)	0. 773	0. 815	0. 828	0. 838	0. 833	0. 833
Our(four)	0. 871	0. 865	0. 875	0. 938	0. 883	0. 900

a the results in [7]

b the results in [18]

Table 2 While L=20,the accuracy of k=1,2,...,K, for different M and M1,taking the biggest two, three or four values of f_m,m=1,2,...,17.

k	10	20	30	40	50	60	70	80	MAX	MAX-k
two	0.605	0.616	0.630	0.638	0.645	0.641	0.638	0.637	0.645	50
three	0.728	0.758	0.769	0.777	0.779	0.780	0.776	0.775	0.782	57
four	0.805	0.840	0.848	0.846	0.843	0.838	0.843	0.845	0.851	27

Sometimes, sequence similarity based approaches are often inadequate in the absence of similar sequences or when the sequence similarity among known protein sequences is not statistically significant [1,4].Our method avoid this problem by extracting the features from sequences based on physiochemical properties.

In our method, it is critical to classify residues based on physiochemical properties of proteins. It's based on four important properties as Cai suggested that amino acid composition, hydrophobicity, polarity [20], and charge properties play more critical roles than other features [4]. If this method is based on one kind of physiochemical property only, it will contain less valuable information, the accuracy will be lower. But if the amino acids are divided into too many classes, the number of sequence features will be too many, and some features will be redundant, so we divide them into five classes.

L is also very important. If it is too small, the number of features will not be enough for prediction, and if it is too large, it will be redundant. The chosen biggest number and the number of nearest neighbors (K) are essential. In this work, while K approaches 30, the accuracy reaches the maximum value.

5 Conclusions

In this letter, we proposed a new method of extracting features from protein sequences based on four physicochemical properties of amino acids classes. And we adopted the k-nearest neighbors' algorithm (KNN) to predict the protein functions. The experimental results show that our method is better than some existing methods. Our method has obtained the local and global information of sequences, so it avoids some problems of sequence similarity based methods. Our method can be used to predict protein function while just knowing the sequences, because it does not require any information other than the primitive sequences.

Acknowledgment

This work is supported by the National Nature Science Foundation of China (Grant 60973082), the National Nature Science Foundation of Hunan province (Grant 07JJ5080), the Planned Science and Technology Project of Hunan Province (Grant 2009FJ3195) and the National Nature Science Foundation of Zhejiang province (Grant Y1090264).

References

[1] L. Y. Han, C. Z. Cai, Z. L. Ji, Z.W. Cao, J. Cui, Y. Z. Chen, Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach, *Nucleic Acids Res*. **32** (2004) 6437–6444.

[2] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool, *J. Mol. Biol.* **215** (1990) 403–410.

[3] W. R. Pearson, D. J. Lipman, Improved tools for biological sequence comparison, *Proc. Natl. Acad. Sci. USA* **85** (1988) 2444–2448.

[4] B. J. Lee, M. S. Shin, Y. J. Oh, H. S. Oh, K. H. Ryu, Identification of protein functions using a machine-learning approach based on sequence-derived properties, *Proteome Sci.* **7** (2009) #27.

[5] T. Kawabata, MATRAS: A program for protein 3D structure comparison, *Nucleic Acids Res.* **31** (2003) 3367–3369.

[6] I. Eidhammer, I. Jonassen, W. R. Taylor, Structure comparison and structure patterns, *J. Comput. Biol.* **7** (2000) 685–716.

[7] A. Vazquez, A. Flammini, A. Maritan, A. Vespignani, Global protein function prediction from protein-protein interaction networks, *Nat. Biotechnol.* **21** (2003) 697–700.

[8] G. Pugalenthi, K. K. Kumar, P. N. Suganthan, R. Gangal, Identification of catalytic residues from protein structure using support vector machine with sequence and structural features, *Biochem. Bioph. Res. Co.* **367** (2008) 630–634.

[9] Y. C. Chen, Y. S. Lin, C. J. Lin, J. K. Hwang, Prediction of the bonding states of cysteines using the support vector machines based on multiple feature vectors and cysteine state sequences, *Proteins* **55** (2004) 1036–1042.

[10] L. J. Jensen, R. Gupta, N. Blom, D. Devos, J. Tamames, C. Kesmir, H. Nielsen, H. H. Staerfeldt, K. Rapacki, C. Workman, C. A. Andersen, S. Knudsen, A. Krogh, A. Valencia, S. Brunak, Prediction of human protein function from post-translational modifications and localization features, *J. Mol. Biol.* **319** (2002) 1257–1265.

[11] L. J. Jensen, M. Skovgaard, S. Brunak, Prediction of novel archaeal enzymes from

sequence-derived features, *Protein Sci.* **11** (2002) 2894–2898.

[12] C. Pasquier, V. J. Promponas, S. J. Hamodrakas, PRED-CLASS: cascading neural networks for generalized protein classification and genome-wide applications, *Proteins* **44** (2001) 361–369.

[13] L. C. Borro, S. R. M. Oliveira, M. E. B. Yamagishi, A. L. Mancini, J. G. Jardine, I. Mazoni, E. H. dos Santos, R. H. Higa, P. R. Kuser, G. Neshich, Predicting enzyme class from protein structure using Bayesian classification, *Genet. Mol. Res.* **5** (2006) 193–202.

[14] I. Halperin, D. S. Glazer, S. Wu, R. B. B. Altman, The FEATURE framework for protein function annotation: modelling new functions, improving performance, and extending to novel applications, *BMC Genomics* **9** (2007) #S2.

[15] J. K. Kim, S. Y. Bang, S. J. Choi, Sequence-driven features for prediction of subcellular localization of proteins, *Pattern Recogn*. **39** (2006) 2301–2311.

[16] Q. B. Gao, Z. Z. Wang, C. Yan, Y. H. Du, Prediction of protein subcellular location using a combined feature of sequence, *Febs Lett*. **579** (2005) 3444–3448.

[17] P. F. Du, T. He, Y. D. Li, Prediction of C-to-U RNA editing sites in higher plant mitochondria using only nucleotide sequence features, *Biochem. Bioph. Res. Co.* **358** (2007) 336–341.

[18] X. Li, B. Liao, Y. Shu, Q. G. Zeng, J. W. Luo, Protein functional class prediction using global encoding of amino acid sequence, *J. Theor. Biol.* **261** (2009) 290–293.

[19] B. Liao, Q. Liu, Q. Zeng, J. Luo, G. Yue, An approach for data selection of protein function prediction, *MATCH Commun. Math. Comput. Chem.* **65** (2011) 459–468.

[20] C. Z. Cai, W. L. Wang, L. Z. Sun, Y. Z. Chen, Protein function classification via support vector machine approach, *Math. Biosci.* **185** (2003) 111–122.