MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **65** (2011) 459-468

ISSN 0340 - 6253

An Approach for Data Selection of Protein Function Prediction

Bo Liao[1*], Qinfeng Liu[1*], Qingguang Zeng[1], Jiawei Luo[1], Guanxue Yue[2]

[1] *School of computer and communication, Hunan University, Changsha Hunan, 410082, China*

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

(Received March 24, 2010)

Abstract: Predicting the function of an unknown protein is an essential goal in bioinformatics. Many methods have been provided to predict the functions of proteins based on sequence similarity. However, they are often inadequate in the absence of similar sequences or when the sequence similarity among known protein sequences is statistically weak. This study aimed to choose some nearest samples dataset at length for identifying protein function, irrespective of sequence and structural similarities. The results of our experiment show that our method is efficient.

1 Introduction

Genome sequencing projects continue to produce unprecedented amounts of novel protein sequence information and large-scale experimental efforts are underway to determine the function of the newly discovered proteins[1], so protein function class prediction is very important and indispensable. The method of determining the function of proteins by experiments is so costly and time-consuming. So, the research of computational approaches to predict functions is important and essential. At present, there are many methods already

* Corresponding author. Fax: +86 731 8821715
E-mail address: dragonbw @163.com (B. Liao)
* E-mail address: dragonbw @126.com (Q.F Liu)

adopt to predict protein function class. There are mainly three different classification methods. The first method based on sequence similarity, such as BLAST[2], Data Mining[3], Neural Network approach [4] and so on. This model is to find the similar sequence by extracting sequence feature in known protein sequence. The second method based on protein-protein interaction, such as GO(Global Optimization) annotation means[5-7] and MRF(Markov random field)[8-9]. This predictive method relies on the number of genome and accuracy of protein-protein interaction database. The third method based on the structure similarity, Cai[10], Kawabata T.[11] and Eidhammer I.[12] to predict function class grounded on structure similarity. This method was proposed not long ago based on three structures, but three structures determined are very difficult, so it is not used widely at present. In order to predict function with great accuracy, we present a method to choose some nearest samples dataset at length for identifying protein function, irrespective of sequence and structural similarities.

2 Dataset

To confirm the relation, we loaded down the 1818 proteins of yeast from ftp://ftpmips.gsf.de/yeast/, which are used to predict protein function, except for eight proteins that cannot be got in the database or whose sequence length is too short. There are 1377 proteins which are known among them, the seventeen functional categories of all proteins were presented in Table 1. In order to discuss the relation better, we choose 1377 protein known for experiment.

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, transport facilities and transport routes	331
Protein synthesis	98	Regulation of metabolism and protein	115

		function	
Interaction with the environment	172	Cellular communication/signal transduction mechanism	110
Cell fate	143	Cell rescue, defense and virulence	201
Biogenesis of cellular components	324	Transposable elements, viral and plasmid proteins	5
Transcription	427		

3 Methods

At first, we reorder all protein sequence of 1377 conforming to the order from the short to the long. In most cases, the gap of the length between two adjacent sequences is very narrow. We set that m is consecutive sequential set. For example, When m=50, we can randomly get a sequence set(e.g. 11-60) for all sequence set(1-1377). We transform amino acid sequence set to profile coding dataset. Profile coding[13][14] is that every amino acid appears in total sequence at some frequency. For example, it is clear that amino acid Q appears at the rate of 1 in Sequence A (QQQQQ) at the first line; for another example, frequency of amino acid G is 0.4, because there are two G in the sequence B(GGTHH); and so on. S_{eq} is the name of amino acid sequence, A_m are concrete amino acid set. Profile encoding of Amino acid sequence (ABCDEFGH) are all lines they correspond to. For instance, the profile coding of sequence A is 00000000001000000000, and encoding of B is 00000000.4000000.200000.40. It's widely agreed that this encoding includes lots of information of evolution [13]. We convert all protein sequence to profile coding, which represent the feature of amino sequence.

Table 2 Profile encoding

S_{eq}	A_m					A	C	I	L	M	V	F	H	W	Y	N	Q	S	T	K	R	D	E	G	P
A	Q	Q	Q	Q	Q	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
B	G	G	T	H	H	0	0	0	0	0	0	0	0.4	0	0	0	0	0	0.2	0	0	0	0	0.4	0
C	K	K	K	K	S	0	0	0	0	0	0	0	0	0	0	0	0	0.2	0	0.8	0	0	0	0	0
D	A	M	M	M	M	0.2	0	0	0	0.8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	D	D	D	F	F	0	0	0	0	0	0	0.4	0	0	0	0	0	0	0	0	0	0.6	0	0	0

F	C	C	C	C	C	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	L	K	M	N	A	0.2	0	0	0.2	0.2	0	0	0	0	0	0.2	0	0	0	0.2	0	0	0	0
H	R	R	R	D	D	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.6	0.4	0	0
I	S	S	T	T	T	0	0	0	0	0	0	0	0	0	0	0	0	0.4	0.6	0	0	0	0	0

Then we choose many kinds of sequence set for test using NNA (Nearest Neighbor Algorithm), NNA[15] can be used to distribute categories of the protein are unknown. For classification problem, suppose there are N_i samples $x_j^{(i)}$ (j=1,2,…, N_i) at class ω_i(i=1,2,…,c). Specification of categories as follow: for a sequence feature vector x, we separately compute the distance between x and N=$\sum_{i=1}^{c} N_i$ samples $x_j^{(i)}$ that their function are known, x will belong to the class of $x_j^{(i)}$ that their distance is lowest. Using this classification measure, judging function of class ω_i is

$$d_i(x) = \min_{j=1,2,...,N_i} \| x - x_j^{(i)} \|, \text{ i=1,2,…,c}$$

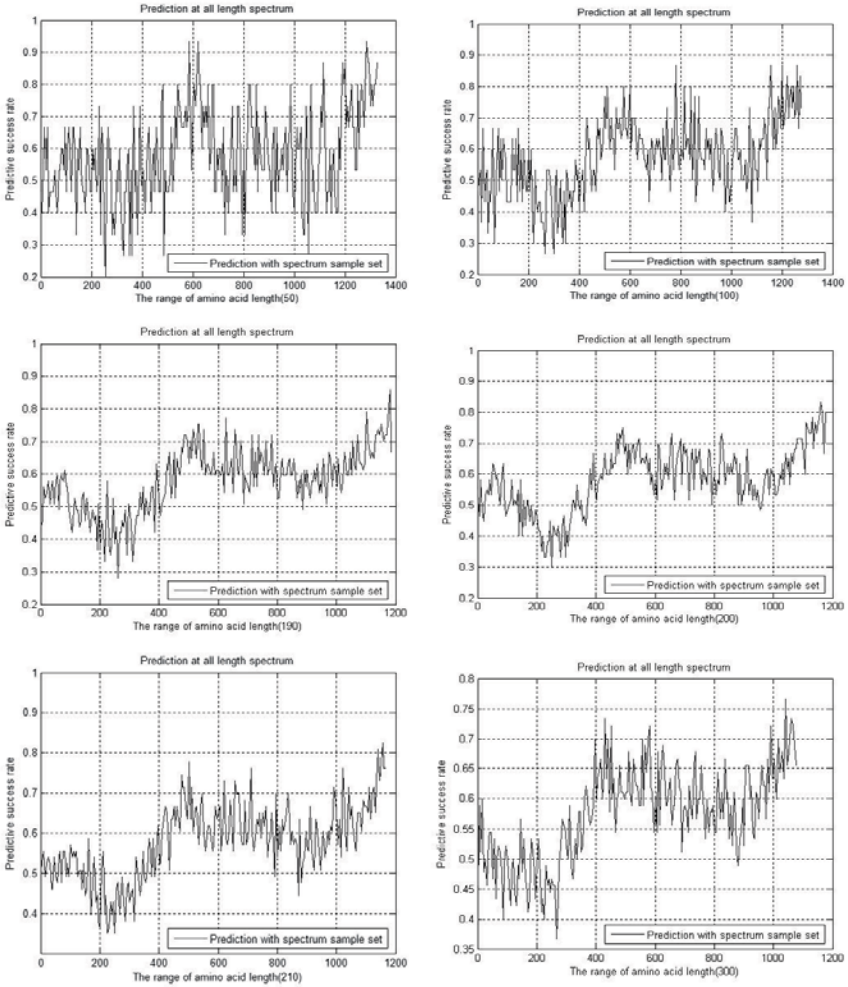
Here $$\left\| x - x_j^{(i)} \right\| = \sum_{i=1}^{c} \frac{\left| x - x_j^{(i)} \right|}{\left| x + x_j^{(i)} \right|}, x, x_j^{(i)} \geq 0, x + x_j^{(i)} \neq 0$$

If $$d_m(x) = \min_{i=1,2,...,c} d_i(x)$$

Then judge $x \in \omega_m$

In other words, x should belong to the m-th class. In all sequence set we select some of them from the first one to the last one at the step length of 5, so we will get 1, 6, 11, 16, 21,…, 1376 picked point. At all picked points, we will take a serious of consecutive sequence set (m). Take example for m=200, the picked point are 1,6,11,16,21,…,1176(1376-200).At those points, sequence set we get separately are 1-200,6-205,11-210,16-215,21-220,…,1176-1375. 30% of m that we randomly get be regarded as test dataset, the remainder should be regarded as train dataset, at all picked points their predictive success rate can be got (See the fourth chart in Fig1). Intention, we select the nearest sequence according to the length of test

sequence set. In the same way, we consider on m=50,60,70,…,1000, because if m is too small, the train dataset is too small, randomness will be larger; if m is too larger, the number of the picked point is small (See the first, last chart in Fig1). In this way, prediction charts are shown base on the value of m. As the layout of paper is limited, we only enumerate some charts, which are very representative in Fig 1. It displays protein function prediction in the continuous amino acid sequence set from the short to the long.



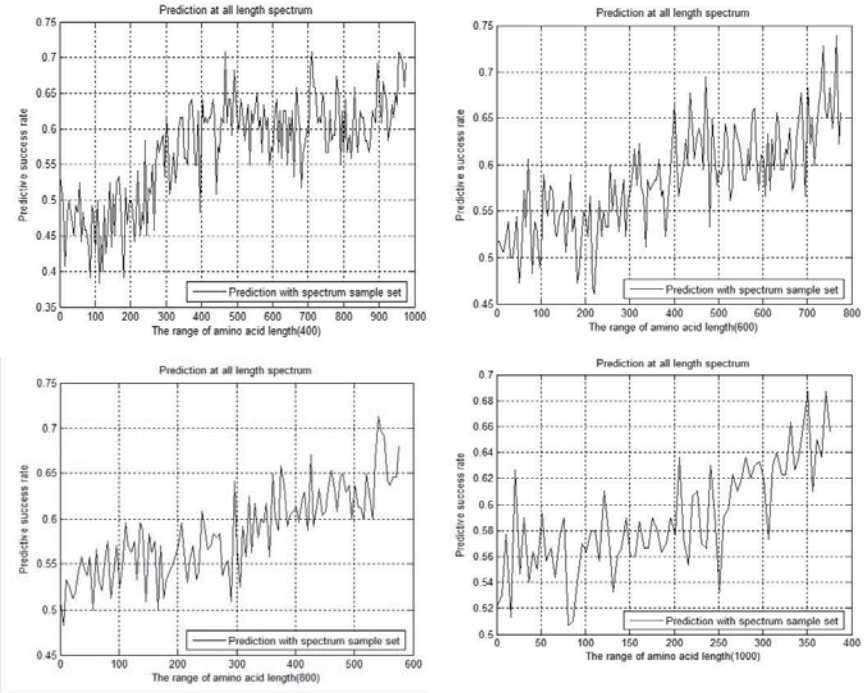


Fig 1 In these chart, In order to depict how to select samples in all sequence set, we choose consecutive sequential set at separate picked point as above, 30% of them are unknown sequences set, 70% are samples sequence set. Here, sample sequence set is not nearest with unknown sequence set at length, especially when m is large. So when m is very larger, the curve wave amplitude changes dramatically. And when m is very small, here, sample is small, the curve wave amplitude also changes dramatically. So we should colligate two factor that proper sample and nearest sample, when $m \approx 200$, result of prediction is the best.

In order to explain the sample we select and large sample, we make some comparative experiment. And the train dataset of blue curve is from a small number of consecutive sequence vectors, the train dataset of red curve is from almost total sequence vectors. We also take m=200 as an example, the first test is the same as the above, 30% of m be regarded as test dataset, the remainder are train dataset. But at the second curve the test dataset is the same as at the first test, only the train dataset is from all total sequence set except for some sequence set at test dataset (See Chart 1 in Fig1). When m=400, 600, 800, 1000, various of comparative charts are depicted as follow in Fig 2.

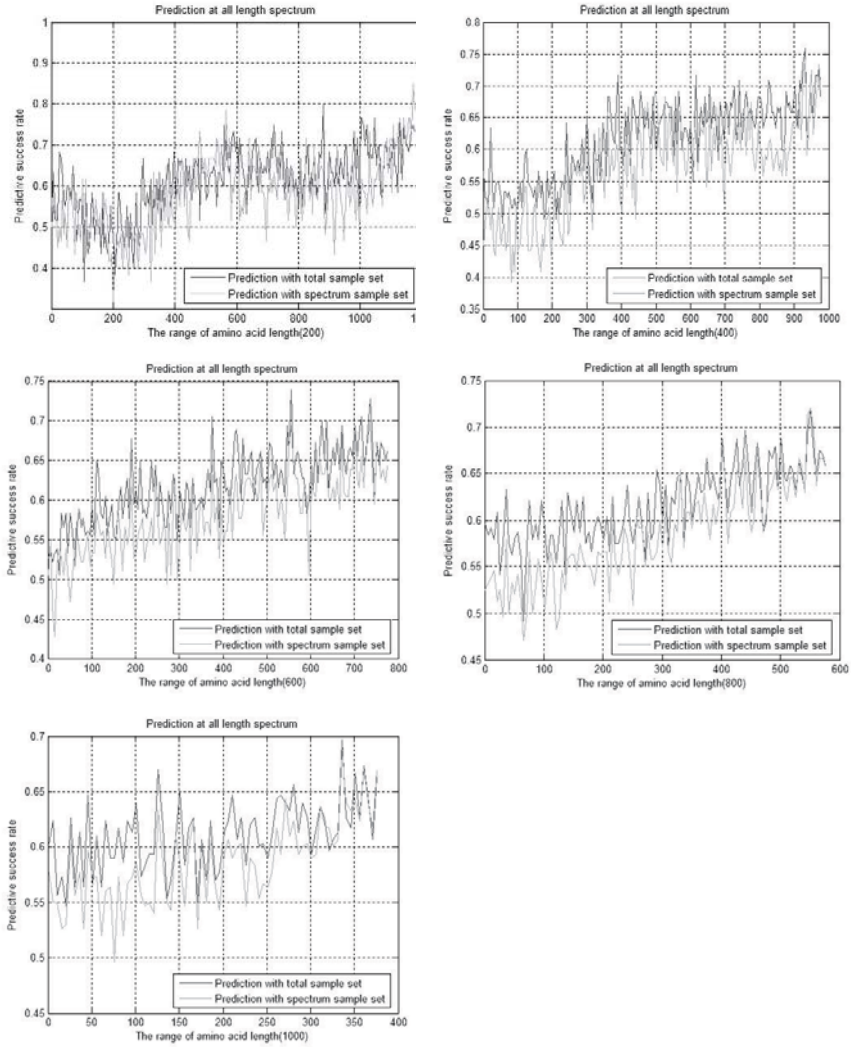


Fig2 These charts show several comparative curves. The red curve is depicted for prediction in small samples, the blue one in the remaindering total sample. As m increases, the samples which we select are far away from the nearest sample, so the blue curve a little higher than the red one when $m \geq 400$. When m=200, the sample at our experiment is very close to the nearest sample, the two curves almost coincide. The chars argue that the method of selecting nearest samples is the best.

3.3 Result and discussion

From the first group graph, at every chart we find that predictive accuracy gradually grows up with sequence length increasing in general. It reflects laws of biological evolution that the shorter the sequence, the greater the mutation rate is. When m is small, all test sequence length is nearly the same long, the number of train dataset is little, so predictive randomness is large; when m is larger, Randomness of the sample selection is too large, predictive rate changes dramatically. The selection of the sequence set (m) should be moderate. Here, m=200, curve wave amplitude changes very little. Its sample is not small and randomness of the sample selection is not too large. It includes two advantages. Therefore, we choose more sample if the number of test sequence set is very small. In other side, we should get the nearest samples for prediction at length.

From the second group comparative graph, when m=200, predictive curve with spectrum sample set, which is a small number of consecutive sequence vectors, approximately coincides with the curve with the remaining total sample set. When m =1000, predictive success rate of the total curve with larger sample is just higher than that curve with little sample. In addition, in other case of m, the difference of predictive success rate between two curves is not large. These experiments clearly show that a good predictive result can be got if we choose the sequence sample as train dataset according to the nearest samples of unknown sequence set at length. When the number of test dataset, which is not very small or not very large, is moderate, for example, the number is 60, m=200, it reflect two advantage the result with small train dataset is as good as the one with the larger.

Conclusion

Above all, from the first group experiment, the longer the sequence length, the better predictive accuracy rate is. For unknown protein sequence set, base on length of test sequence set, we should select the nearest samples of test sequence set, predictive success rate of protein sequence is good. Usually, the samples are twice as large as the unknown sequence set. If the unknown sequence set is very small, the sample should have a determinate number. For instance, the number should be more or equal 200. Therefore, if we want to know the

function of unknown protein sequence quickly, we can use this method to choose train dataset to save time. If the train dataset is very large, we do not need choose all samples but some necessary. Next , this method will be used on other species for further prediction and get better predictive rate with little sample set.

Acknowledgment

This work is supported by the National Nature Science Foundation of China (Grant 60973082, 60873184), the National Nature Science Foundation of Hunan province (Grant 07JJ5080, 06JJ2090), the Planned Science and Technology Project of Hunan Province (Grant 2009FJ3195) and the National Nature Science Foundation of Zhejiang province (Grant Y1090264).

References

[1] A. Al-Shahib, R. Breitling, D. R. Gilbert, Predicting protein function by machine learning on amino acid sequences–a critical evaluation, *BMC Genomics* **8** (2007) 78, 1–10.

[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] R. D. King, A. Karwath, A. Clare, L. Dehaspe, The utility of different representations of protein sequence for predicting functional class, *Bioinformatics* **17** (2001) 445–454.

[4] L. J. Jensen, R. Gupta, H. H. Staerfeldt, S. Brunak, Prediction of human protein function according to Gene Ontology categories, *Bioinformatics* **19** (2003) 635–642.

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

[6] M. Zhu, L. Gao, Z. Guo, Y. Li, D. Wang, J. Wang, C. Wang, Globally predicting protein functions based on co–expressed protein-protein interaction networks and ontology taxonomy similarities, *Gene* **391** (2007) 113–119.

[7] Y. R. Cho, A. Zhang, Predicting protein function by frequent functional association pattern mining in protein interaction networks, *IEEE Trans. Inf. Technol. Biomed.* **14** (2010) 30–36.

[8] M. Deng, K. Z. Shipra Mehta, T. Chen, F. Sun, Global protein function predition from protein-protein interaction data, *Comp. Biotech.* **10**(2003) 947–960.

[9] M. Deng, T. Chen, F. Sun, An integrated probabilistic model for functional prediction of proteins, *J. Comput. Biol.* **11** (2004) 463–475.

[10] Y. D. Cai, A. J. Doig, Prediction of Saccharomyces cerevisiae protein functional class from functional domain composition, *Bioinformatics* **20** (2004) 1292–1300.

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

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

[13] H. Zhu, L.Yoshihara, K. Yamamori, Prediction of protein secondary structure by multi–modal neuralnetworks, *Proc. International Joint Conference on Neural Networks(UCNN'02)*, 2002, pp. 280–285.

[14] H. Bohr, J. Bohr, S. Brunak, R. M. Cotterill, B. Lautrup, L. Norskov, O. H. Olsen, S. B. Petersen, Protein secondary structure and homology by neural networks. The alpha-helices in rhodopsin, *FEBS Lett.* **241** (1988) 223–228.

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