A Novel Protein Motif Finding Algorithm for Classification of the Ligase Subfamilies

Deng-Kuan Sun[1], Tong-Liang Zhang[1], and Yong-Sheng Ding[1,2*]

1) College of Information Sciences and Technology
2) Engineering Research Center of Digitized Textile & Fashion Technology, Ministry of Education
Donghua University, Shanghai 201620, P. R. China
*Corresponding to: ysding@dhu.edu.cn

Abstract. The algorithm of extracting motifs from a family or subfamily is still a hot spot in bioinformatics. It not only contributes to understand functions of proteins and predicts the classification which a unknown protein sequence belongs to, but also helps to study the protein-protein interaction. In this paper, we present a novel algorithm to extract motifs of a subfamily, which is based on feature selection and position connection. Position connection is applied to generate motifs, which is the hybrid method with mechanism of vote decision-making to construct the classifier of the ligase subfamilies. Through testing in the database, more than 95.87% predictive accuracy is achieved. The result demonstrates that this novel method is practical. In addition, the method illuminates that motifs play an important role to classify proteins and research the characteristics of the subfamilies or families of protein database.

Keywords: motifs extracting, protein classification, vote decision-making, ligase enzyme

1 Introduction

Protein sequence is composed of 20 kinds of amino acids. The order of amino acids in sequence determines the spatial structure and function. In protein sequence, some important segments, which are called motif, exist in almost every protein sequence in protein family or sub-family. The motifs in protein are able to represent the function of protein. Therefore, it is significant to find the important motifs in protein sequence in the solution of what family and subfamily a protein belongs to.

Extracting the important motifs in protein family is a NP (nondeterministic polynomial time) hard problem. Many methods have been developed to find motifs in protein sequence and DNA sequence. However, until now there is not a perfect algorithm which can extract motifs from some protein families accurately. There are mainly two kinds of algorithms for motif finding in the existing works. One is finding some small or short motifs which are based on strings analyses and probability methods. These algorithms include MEME algorithm [1] and Gibbs algorithm [2, 3]. The other is heuristic algorithm, which can extract some larger motifs. The method was summarized by Wasserman and Krivan, and it is represented by some artificial

M. Chetty, S. Ahmad, A. Ngom, S. W. Teng (Eds.): PRIB2008,
Supplementary Conf. Proc., pp.49-60, 2008.

intelligence such as GA (Genetic Algorithm) and RBF (Radial Basis Function). Nevertheless, the method has some shortcomings such as the trend getting into local optimization, larger compute complexity. Some useful improvements have been established, such as AlignACE [4], BioProspector [5], MDScan [6], MEME [1,7], and MotifSampler [2]. An integrated algorithm has also been introduced based on vote tactic. Hon and Jain exploit MaMF algorithm [15], which can be used to extract motifs of mammal. Based on some feature selection and connecting approach, where the features are similar to *A-x-B* (*A*, *B* is one of the amino-acids, and *x* is the spaces between *A* and *B*), Chang et al. proposed a motif finding algorithm and used it for extracting the motifs of C2H2 Zinc Finger and EGF proteins [8]. However, as for those approaches there are some shortcomings in compute complexity [2, 9].

In this study, a novel motif finding algorithm is proposed. Compared with previous works, the proposed algorithm has faster speed and could get longer motif in protein family. The novel algorithm is used to extract motifs of ligase subfamily. According to the results of motifs in ligase subfamily, protein sequence can be determined what ligase subfamily it belongs to. The satisfying results indicate that the motif finding algorithm is effective. The potential useful tools for predicting protein family could be developed based on the proposed algorithm. The sections are organized as following: In section 2, we introduce the motif finding algorithm. The ligase subfamily prediction based on motifs is introduced in section 3. The discussion of this algorithm is explained in section 4. Finally, the conclusion is listed in section 5.

2 Motif finding algorithm

A motif is hided in all of the sequences of the protein family, which is an amino acid segment with different spaces. The problem of finding useful motifs in protein family is a NP hard one. Even if the heuristic algorithms can solve the problem, but they are time-consuming. Chang et al. [8, 14] have presented a useful method to search motif through constructing motif features. Inspired by prior work, we develop a novel algorithm for finding useful motifs in protein family. According to the idea of Chang's work, important feature selection in protein family is the first step in our algorithm. Then, linking the important features based on its position in sequences. Meanwhile, combining the mechanism of vote decision-making, a prediction system is constructed to identity the protein family of a sequence. The algorithm is described as below:

1) Feature selection, which is used to bring some important features;
2) Linking important features, which is used to bring candidate motifs;
3) Selecting motifs, which is used to bring a motif bag;
4) Constructing prediction system, which is used to determine the protein family of a sequence.

2.1 Feature selection

According to the motif format of Prosite Database [10], *AB-x(5)-CD-x(1,3)-EF* is generic, where *A*, *B*, *C*, *D*, *E* and *F* are one of twenty amino acids, *x(5)* is the gap

between *B* and *C*, and *x(1,3)* shows that the gap is any number between 1 and 3. In [8], the motif can be divided into the following formats:

A-x(0)-B,B-x(5)-C,C-x(0)-D,D-x(1,3)-E,E-x(0)-F.

These formats can be summarized into the pattern: A_1-$I_{1,2}$-A_2, where A_1 and A_2 is one of twenty amino acids, and $I_{1,2}$ is the gap between A_1 and A_2. The statistical results in Prosite Database [10] indicate that $I_{1,2}$ is less than twenty in more than 99% of motifs. Thus, there are total 8,000 candidate features A_1-$I_{1,2}$-A_2 ($20\times20\times20$). Computing the *PAV* value (that is the ratio in which candidate feature appears in a subfamily) of all of candidate features in any subfamily, some important features are selected out. The format is displayed the following:

$$PAV_i = \frac{\sum_{j=1}^{n_c} f_j}{n_c},\ (i = 1,...,8400) \tag{1}$$

Where PAV_i represents the *PAV* value of any candidate feature; f_i is a 0-1 function, determined by Eq. (2), and n_c is the number of protein sequences in protein family.

$$f_i = \begin{cases} 1 & A_1\text{-}I_{1,2}\text{-}A_2 \subseteq S \\ 0 & otherwise \end{cases} \tag{2}$$

where *S* is the protein sequence. As an example, the meaning of *PAV* is displayed in the Fig.1.

Seq_1: MVTSAGVGHAEYNNGADVQHADYAHLTSVGQVEQKPLGGRL
Seq_2: MPESTQQSHLSLDHEKMQQPPKGFTERSKTKPNLADFETYQKLY
Seq_3: MSPSAIAEKKQVDNIQEIKDKNQEPAHHEYEHLTNVGVVKQKPI
Seq_4: MTPHSTHVTLDHEGIIQPPAEFKERSKSKPNLADFETYSKMYKESI
Seq_5: MSPAVDTASTAKDPISVMKSNASAAAADQIKTHEYEHLTSVPIVQ
Seq_6: MSPSAVQSSKLEEQSSEIDKLKAKMSQSAATAQRKKEHEYEHLT

Fig.1. It is set that there are six protein sequences (that is, n_c is six) in a family. A_1-$I_{1,2}$-A_2 can be represented as MV,M-T,M--S, e,g. S represents any of these six protein sequences. In the Seq_1, the value of f_i of MV is 1 because it belongs to this one, while the value of f_i of MP is 0 because it does not. Although the number of QQ in Seq_2 is two, the value of f_i is 1. Therefore, the PAV_i value of M-P is 4/6, and one of AD is 4/6, and one of PP is 2/6, and so on.

Threshold is set to select important features. Here, we set $r_0=0.95$. It means the important feature occurrence is more than 95% in sequences of protein family.

Through feature selection, some important features, which have the characteristic of motif, are obtained. However, it can be found that different subfamilies have some same features, and if these features are constructed into a classifier, they will affect the result. Therefore, it is necessary to link these features into a series of long and complex features, which could contribute to advance the prediction result based on the fact that amino acids of motif appear in certain order.

In the work of Chang et al. [8], a link method was given. Although it is effective, the method has big computing complex. For example, it is assumed that there are five features including *C-x(10)-D*, *C-x(11)-E*, *C-x(12)-F*, *C-x(13)-G*, *C-x(14)-H*. the connection times in the method of Chang et al. [8] are more than 100,000. In this

study, we propose a feature connection method based on its position in sequences which could decrease the link times distinctly. That is, if the above same features are given and their head “C” appears in the same position of a sequence, they can be connected into *C-x(10)-DEFGH* by only once. In fact, this thought represents the characteristic of motifs in biology sequence. In the process of evolution, the motif changes less and amino acids in it have certain positives. Sequence evolution is the process which includes insertion and cutting off father sequences.

The feature connection algorithm based on the position in sequence is described in the following steps:

1) 10% of all of the sequences of a subfamily are chosen as a training set, and the rest are testing set;

2) These important features are labeled their positions in the training set.

3) These features are connected based on that they have the same head position, or the same terminal, or one’s head and another’s terminal is the same;

4) The *PAV* value of these connected features is tested, and these features, of which the *PAV* is more than 0.9, are considered as candidate motifs.

And the process is displayed in the Fig.2.

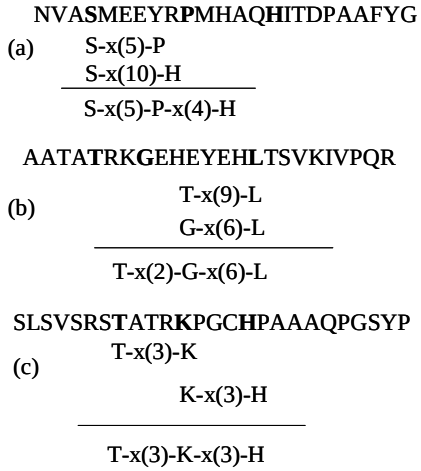

Fig.2. The process of connection feature based on the position in sequence. (a) The heads of two features are the same; (b) The terminals of two features are the same; (c) The head of one feature and the terminal of another feature are the same

2.2 Motifs generation

Through feature selection and feature link, some candidate motifs, whose format is $A\text{-}I_{1,2}\text{-}B\text{-}I_{2,3}\text{-}C$, are obtained. However, through testing the classification result of these candidate motifs, the true positive is also low. Therefore, in order to improve the result further, these candidate motifs must be connected further.

It is proposed that candidate motif set of a subfamily is $M=\{M_1,M_2,\ldots\ldots,M_n\}$. There are n elements in the set, for instance, $A\text{-}I_{1,2}\text{-}B\text{-}I_{2,3}\text{-}C$, has two elements: $A\text{-}I_{1,2}\text{-}B$ and $B\text{-}I_{2,3}\text{-}C$. Therefore, the step of motif generation is displayed in the following:

1) Initialize all of the elements in order that they are not labeled;

2) Choose a candidate motif from the above set randomly;

3) Connect these candidate motifs which have some same elements and test in the training set by two criterions:

(1) It can identify all of the sequences of its subfamily largely;

(2) It can not classify the sequences of other subfamily into its subfamily.

If these criterions are not met, the link is cancelled. Then, if it is connected, the connected candidate motif is considered as a motif and is labeled. For example, $A\text{-}I_{1,2}\text{-}B\text{-}I_{2,3}\text{-}C$ is a motif. If it belongs to some subfamily, the *PAV* of $A\text{-}I_{1,2}\text{-}B$ and $B\text{-}I_{2,3}\text{-}C$ would be high. If it does not belong to this subfamily, their *PAV* would be low. Therefore under the first condition, their *PAV* is considered as 0.98 and 0.97, then after they are connected, the *PAV* of $A\text{-}I_{1,2}\text{-}B\text{-}I_{2,3}\text{-}C$ would be 0.97. Yet under the second condition, their *PAV* is considered as 0.08 and 0.06, then after they are connected, the *PAV* of $A\text{-}I_{1,2}\text{-}B\text{-}I_{2,3}\text{-}C$ could be less than 0.0048. As a result, through two criterions, some motif features can be obtained.

4) When the candidate motif attempts to connect with other candidate motifs of a subfamily, the second will be done again until all of motifs are labeled.

For instance, set *M* includes *Y-x(1)-G-x(10)-G*, *Y-x(1)-G-x(1)-D*, *G-x(1)-D-x(4)-S*, *Y-x(3)-D-x(8)-G*. The process is displayed by the following steps:

(1) *Y-x(1)-G-x(10)-G* is chosen as a motif, and it has the same element with *Y-x(1)-G-x(1)-D*. After they are connected, a new motif *Y-x(1)-G-x(1)-D-x(8)-G* is obtained;

(2) If the new motif meets the two criterions and has the same element with *G-x(1)-D-x(4)-S*. After they are connected, the new motif is renewed as *Y-x(1)-G-x(1)-D-x(4)-S-x(3)-G*.

(3) If *Y-x(1)-G-x(1)-D-x(4)-S-x(3)-G* meets the criterions, connecting it and *Y-x(3)-D-x(8)-G*, the terminal motif is obtained and tested. If it meets the criterions, it is the motif of the subfamily.

Through these steps, we will find some motifs of a protein family. And it contributes to the following classifier construction.

2.3 The mechanism of vote decision-making

From the first to third steps, motifs of a subfamily are obtained. When any of them is used to classify protein sequences into their subfamily, it is found that the prediction result is also not good since they have some common functions, which belongs to the same family (that is, there are the same amino acid groups). Therefore, the mechanism of vote decision-making is introduced after these motifs of a subfamily are constructed into a motif bag. When a queried sequence is given, if it has a given rate of motifs in the motif bag, it will be considered as one of the subfamily. The process is displayed in the Fig.3.

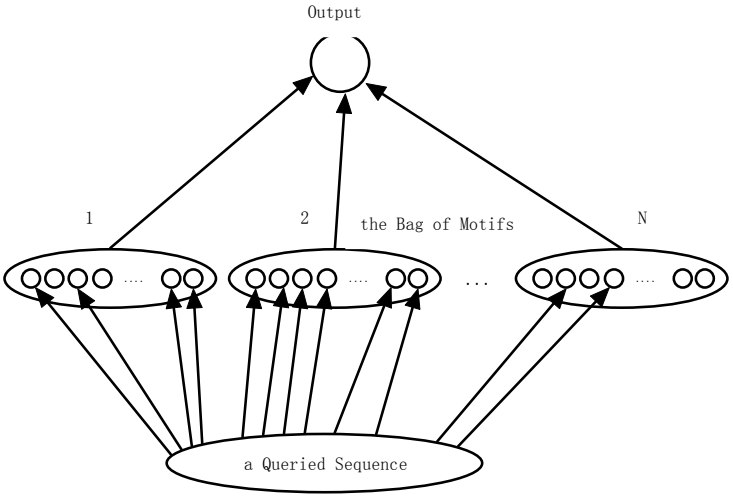

Fig. 3. The process of the classifier been implemented. Firstly, the queried sequence is recognized by the bag of motifs of the subfamily. Secondly, the number of the element of the bag of motifs belongs to the sequence is computed. Thirdly, the ratio of the elements which the sequence includes to the size of the bag is compared, and the classifier outputs the subfamily whose ratio is the most.

3 Application in classification ligase subfamily

Ligase is one kind of enzyme. The function of ligase is connecting chemistry bond in some environment. The EC label in enzyme database [10] is E6. Several approaches have been proposed for predicting enzyme family and subfamily. Chou and Elord [11] developed covariant discriminant algorithm for classification oxidoreductases. Chou and Cai predicted 6 types ligase subfamily, 94.89% accuracy is obtained by jackknife test [12]. Huang et al. developed AFK-NN algorithm for predicting six enzyme families and oxidoreductases subfamily [13]. Here, we develop a voting classification system for predicting ligase subfamily.

3.1 Dataset

Ligase database coming from [10] is used to measure the algorithm of extracting motifs. In the database, all of enzymes have a label of EC. As for these subfamilies and sequences, which do not meet the need of the possibility, such as some enzymes that belong to several classes and some subfamilies that only have several or tens of sequences, they are not chosen. Therefore, in the paper, all of databases are presented in the Table I.

Table I. The ligases databased used in the algorithm

Classws	Number	Name of the Subfamilies
Class 1	79	ATP-dependent AMP-binding enzyme family
Class 2	177	class-I aminoacyl-tRNA synthetase family. TyrS type 1 subfamily
Class 3	204	class-II aminoacyl-tRNA synthetase family
Class 4	215	class-I aminoacyl-tRNA synthetase family. ValS type 1 subfamily
Class 5	108	class-I aminoacyl-tRNA synthetase family. MetG type 1 subfamily
Class 6	216	class-II aminoacyl-tRNA synthetase family. Type-1 seryl-tRNA synthetase subfamily
Class 7	97	class-II aminoacyl-tRNA synthetase family. ProS type 3 subfamily
Class 8	236	class-II aminoacyl-tRNA synthetase family. Phe-tRNA synthetase alpha chain type 1 subfamily
Class 9	80	succinate malate CoA ligase beta subunit family
Class 10	163	glutamine synthetase family. Type 3 subfamily
Class 11	102	NAD synthetase family
Class 12	51	prokaryotic GSH synthase family
Class 13	148	D-alanine--D-alanine ligase family
Class 14	227	AIR synthase family
Class 15	184	formate--tetrahydrofolate ligase family
Class 16	251	adenylosuccinate synthetase family
Class 17	234	argininosuccinate synthase family. Type 1 subfamily
Class 18	98	GARS family
Class 19	195	FGAMS family
Class 20	151	carA family
Class 21	149	accA family
Total	3365	

3.2 Results

According to the steps of motif finding, the motifs bag of each ligase subfamily is obtained. Through testing the ligase database, some better classification results and motifs of a subfamily are obtained. Taking class 1(ATP-dependent AMP-binding enzyme family) as an example, it is taken as a positive set, and others are integrated into a negative set. Firstly, the *PAV* value is set as 0.95, then 1904 important features are obtained. Secondly, through connecting these features and testing them, 42 candidate motifs are obtained. Thirdly, through connecting candidate motifs further (such steps of connecting have been instructed in section 2.2.), the motif set is obtained and is shown in Fig.4. Fourthly, the motif bag is composed of all elements of the motif set and is used to identify all of the sequences of the subfamily, and the accuracy 98.7% is obtained. With error-control strategy, it can be obtained in section 2.2.

S - x(5) - P - x(4) - H - x(3) - G
G - x(4) - PI - x(7) - G
SG - x(1) - T - x(4) - G
G - x(3) - G - x(4) - YP
D - x(4) - T - x(8) - PL
L - x(3) - R - x(1) - G - x(9) - F
TS - x(3) - G - x(2) - K
A - x(4) - G - x(0) - A - x(5) - G - x(4) - V
T - x(4) - G - x(2) - K - x(3) - H
E - x(2) - L - x(11) - P
LY - x(4) - V - x(7) - H - x(2) - W
L - x(1) - TS - x(2) - T - x(4) - G

Fig.4. Motifs of class 1.

The test results of all of 21 classes of ligase and their motif bag are displayed in Table II and Table III, respectively.

Table II. The classifier result of 21 classes ligase

Ligases families	True positives	False positives
Class 1	98.7%	2.88%
Class 2	97.7%	2.22%
Class 3	96.6%	3.93%
Class 4	99.5%	2.27%
Class 5	96.3%	0.029%
Class 6	92.1%	1.06%
Class 7	92.8%	2.31%
Class 8	91.1%	0.12%
Class 9	95%	3.75%
Class 10	91.4%	1.41%
Class 11	91.2%	1.27%
Class 12	96.1%	2.58%
Class 13	93.2%	1.64%
Class 14	94.3%	0.36%
Class 15	95.6%	1.29%
Class 16	98.8%	2.73%
Class 17	100%	0.68%
Class 18	93.9%	2.28%
Class 19	100%	0.21%
Class 20	96%	1.74%
Class 21	99.3%	3.52%

Table III. Motif bags of 21 classes of ligase

Class 1: ATP-dependent AMP-binding enzyme family		
G-x(4)-PI-x(7)-G	S-x(5)-P-x(4)-H-x(3)-G	G-x(3)-G-x(4)-Y-x(0)-P
SG-x(1)-T-x(4)-G	D-x(4)-T-x(8)-PL	L-x(3)-R-x(1)-G-x(9)-F

TS-x(3)-G-x(2)-K	A-x(4)-GA-x(5)-G-x(4)-V	T-x(4)-G-x(2)-K-x(3)-H
E-x(2)-L-x(11)-P	LY-x(4)-V-x(7)-H-x(2)-W	L-x(1)-TS-x(2)-T-x(4)-G
Class 2: class-I aminoacyl-tRNA synthetase family. TyrS type 1 subfamily		
Y-x(1)-G-x(1)-D-x(4)-S-x(3)-G		K-x(1)-GK
E-x(6)-Q-x(2)-D-x(3)-L	D-x(15)-G-x(1)-DQ	G-x(2)-T-x(3)-GDP
Y-x(1)-G-x(1)-D-x(4)-S-x(1)-H-x(1)-G		
Class 3: class-II aminoacyl-tRNA synthetase family		
R-x(1)-E-x(2)-G-x(3)-G-x(2)-R-x(1)-R-x(5)-D-x(1)-H	H-x(7)-ER-x(6)-E	
Class 4: class-I aminoacyl-tRNA synthetase family. ValS type 1 subfamily		
PP-x(1)-N-x(1)-TG-x(4)-GH-x(7)-D		N-x(7)-G-x(12)-R
Class 5: class-I aminoacyl-tRNA synthetase family. MetG type 1 subfamily		
N-x(8)-N-x(3)-R	D-x(4)-H-x(4)-P	D-x(2)-G-x(6)-A
YVW-x(2)-A-x(3)-Y	N-x(7)-PW	D-x(4)-G-x(1)-CP-x(1)-C
L-x(1)-D-x(4)-G-x(1)-C-x(2)-C		
Class 6: class-II aminoacyl-tRNA synthetase family. Type-1 seryl-tRNA synthetase subfamily		
G-x(4)-G-x(2)-R-x(1)-H-x(1)-F-x(1)-K-x(1)-E		S-x(1)-S-x(5)-Q
R-x(1)-E-x(0)-A-x(3)-G-x(7)-R-x(1)-H-x(1)-F-x(3)-E		G-x(7)-G-x(4)-H-x(1)-F
D-x(1)-E-x(1)-W-x(1)-P	R-x(1)-E-x(14)-H-x(3)-K	P-x(12)-E-x(4)-G-x(7)-R
Class 7: class-II aminoacyl-tRNA synthetase family. ProS type 3 subfamily		
R-x(1)-TSE-x(15)-DL	RPT-x(18)-LP-x(10)-R	Q-x(2)-T-x(1)-H-x(5)-F
R-x(8)-H-x(1)-D-x(2)-G-x(3)-P	P-x(10)-E-x(5)-GF-x(2)-E	RPTSE
Class 8: class-II aminoacyl-tRNA synthetase family. Phe-tRNA synthetase alpha chain type 1 subfamily		
G-x(1)-A-x(3)-G-x(2)-R	W-x(1)-E-x(2)-G-x(1)-G-x(5)-V	N-x(4)-H-x(2)-R-x(3)-D
G-x(15)-F-x(3)-N-x(7)-R-x(3)-D	R-x(3)-F-x(0)-P-x(2)-E	F-x(1)-F-x(1)-EP-x(2)-E-x(1)-D
Class 9: succinate malate CoA ligase beta subunit family		
K-x(1)-Q-x(2)-AGGRGK	F-x(1)-D-x(1)-GG	S-x(2)-G-x(4)-E-x(6)-P
E-x(1)-Y-x(6)-R-x(9)-S-X(2)-G	A-x(1)-F-x(4)-G	G-x(0)-A-x(1)-L-x(5)-D
Class 10: glutamine synthetase family. Type 3 subfamily		
Q-x(1)-E-x(12)-D	NG-x(1)-G-x(1)-H	
Class 11: NAD synthetase family		
S-x(1)-G-x(1)-DS	G-x(2)-T-x(10)-P-x(5)-K	R-x(16)-V-x(2)-T
G-x(9)-G-x(3)-K	G-x(2)-D-x(14)-L	T-x(4)-E-x(6)-T
Class 12: prokaryotic GSH synthase family		
D-x(7)-E-x(1)-N-x(1)-T-x(1)-PT	G-x(8)-I-x(5)-E-x(1)-N-x(1)-T	G-x(1)-D-x(9)-N-x(3)-PT
R-x(2)-PP-x(8)-TT-x(1)-L	N-x(5)-R-x(3)-EK	Q-x(8)-GD-x(1)-R
F-x(9)-K-x(0)-P-x(0)-L-x(4)-G	R-x(7)-R-x(5)-G	D-x(7)-E-x(4)-S
I-x(7)-N-x(1)-T-x(1)-T-x(1)-P		
Class 13: D-alanine--D-alanine ligase family		
HG-x(2)-GE-x(1)-G	N-x(2)-PG-x(4)-S	KP-x(3)-GSS
PG-x(1)-T-x(12)-G		
Class 14: AIR synthase family		
ND-x(4)-GA	M-x(2)-ND-x(4)-GA-x(1)-P-x(4)-DY	HSNG
D-x(1)-V-x(2)-K-x(3)-A-x(11)-D-x(2)-AM		G-x(14)-G-x(2)-L

Class 15: formate--tetrahydrofolate ligase family		
P-x(4)-KG-x(5)-G-x(10)-N	GPF-x(1)-N	GA-x(15)-L
LV-x(5)-T-x(2)-G-x(8)-G	GG-x(2)-FG-x(2)-GG-x(5)-P-x(4)-NL	
R-x(8)-G-x(4)-A-x(1)-G-x(1)-G-x(5)-P-x(10)-D	Q-x(13)-F-x(1)-NIA-x(1)-G-x(2)-S	
Class 16: adenylosuccinate synthetase family		
G-x(3)-GDE-x(1)-KG-x(3)-D-x(13)-GG		G-x(1)-NAGH
Class 17: argininosuccinate synthase family. Type 1 subfamily		
SGG-x(1)-DT	HG-x(1)-T-x(2)-GNDQ-x(1)-RF	A-x(15)-G-x(1)-D
Class 18: GARS family		
GA-x(5)-AE-x(7)-I	I-x(12)-P-x(6)-GE-x(3)-G	R-x(18)-GD
Class 19: FGAMS family		
E-x(1)-H-x(7)-P-x(2)-G	GA-x(2)-G-x(11)-G	
HN-x(1)-P-x(4)-P-x(2)-GA-x(2)-G-x(1)-GG-x(2)-RD-x(4)-G		
Class 20: carA family		
P-x(3)-I-x(0)-C-x(1)-G-x(1)-Q	G-x(2)-E-x(3)-D-x(1)-S-x(3)-Q-x(6)-P	E-x(2)-F-x(1)-T-x(3)-G-x(6)-DPS
G-x(3)-F-x(1)-TD-x(1)-SGY-x(5)-D-x(1)-SY-x(2)-Q		
Class 21: accA family		
G-x(11)-G-x(3)-P-x(1)-G-x(1)-R	G-x(3)-P-x(3)-RK-x(7)-A-x(2)-F	G-x(10)-FG-x(7)-R
GA-x(5)-AE-x(7)-I	I-x(12)-P-x(6)-GE-x(3)-G	R-x(18)-GD
GA-x(5)-A-x(8)-I		

4. Discussion

Through implementing the algorithm in the subfamilies of ligases, the total result of 95.87% is obtained, and this is higher than that of Chou's algorithm (94.89%). Moreover, motifs of such subfamilies are also obtained. Referring to the Prosite database [10], the format of these extracted motifs is accorded with the format of motifs in the database. Therefore, the proposed algorithm is effective in these facts such as finding out some function groups of protein sequences, protein subfamily classification, and so on.

Meanwhile, the algorithm in this study is superior to Chang's in several facts:

1) The complexity of the algorithm is less than Chang's. This can be represented in above introduction.

2) The process of extracting motifs meets the biological significance better. And this is displayed in the process of linking important features based on their position in a sequence.

3) The algorithm is used to the subfamilies of ligases, therefore, the algorithm could have more important sense in some fields, such as drug discovery.

Moreover, there are some differences between the algorithm and Smith's "MOTIF" [14]. For example, in our algorithm, we think the format --- 2-amino acids which be abandoned by Smith as important feature of this algorithm. In the algorithm, through feature selection, some interested features are restricted in small range of 8000 random features. Therefore, computing complexity is reduced. Meanwhile the algorithm is not restricted by gaps.

In addition, although the true positive is good, the false positive is not perfect. Therefore, some methods should be needed so that the result can be advanced. Moreover, although some characteristic of motifs are used in the process of extracting motifs, the biological significance of the algorithm will be enhanced further if the process of evolution, such as insert and cutting off of a sequence, is better expressed.

5 Conclusion

A novel motif finding algorithm is proposed. Motif bag is used to represent the characteristic of protein family. The distinct merits are fast and long motif obtained. In protein family, some units for creating motif are determined by feature selection. The search space is reduced. Feature link based on the position in protein sequence reduces the computation time distinctly. The method of motif connection could produce long motif. The motif finding algorithm is applied to predict ligase subfamily. The motif bags of 21 ligase subfamilies are extracted. The promising results indicate that the motif finding algorithm is effective and practical. In the future work, we will develop a useful web server for predicting enzyme family and subfamily. Moreover, the motif finding algorithm is improved for flexible motif.

Acknowledgments

This work was supported in part by the Key Project of the National Nature Science Foundation of China (No. 60534020), the National Nature Science Foundation of China (No. 60775052), the Cultivation Fund of the Key Scientific and Technical Innovation Project from Ministry of Education of China (No. 706024), International Science Cooperation Foundation of Shanghai (No. 061307041), and Specialized Research Fund for the Doctoral Program of Higher Education from Ministry of Education of China (No. 20060255006).

References

1. Bailey TL, Elkam C.: Fitting a mixture model by expectation maximization to discover motifs in biopolymers[C]. Proc. Of the 2nd International Conf. Int. Sys. Mol. Biol., 28-36. (1994)
2. Thijs G, Marchal K, Lescot M.: A Gibbs sample method to detect over-represented motids in upstream regions of coexpessed genes. Journal of Computational Biology, 9, 447-464. (2001)
3. Lawrence CE, Altschul SF, Bogouski MS: Detecting subtle sequence signals: A Gibbs sample strategy for multiple alignment. Science, 26,2208-214. (1993)
4. Roth FP, Hughes JD, Estep PW, Church GM: Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat Biotechnol, 16, 939-945. (1998)

5. Liu X, Brutlag DL, Liu JS: BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pac Symp Biocomput, 127-138. (2001)
6. Liu XS, Brutlag DL, Liu JS: An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. Nat Biotechnol, 20,835-839. (2002)
7. Bailey TL: Unsupervised learning of multiple motifs in biopolymers using expectation maximization. Machine Learning, 21, 51-80. (1995)
8. Bill C. H. Chang and Saman K. Halgamuge :.Protein motif extraction with neuro-fuzzy optimization. Bioinformatics, 18, 1084-1090. (2002)
9. Tompa M, Li N, Bailey TL, Church GM, De MB, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, Makeev VJ, Mironov AA, Noble WS, Pavesi G, Pesole G, Regnier M, Simonis N, Sinha S, Thijs G, van HJ, Vandenbogaert M, Weng Z, Workman C, Ye C, Zhu Z: Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol, 23, 137-144. (2005)
10. Enzyme nomenclature database, http://www.expasy.org/enzyme
11. Kuo-Chen Chou and David W. Elord.: Prediction of Enzyme Family Classes. Journal of Proteome Research, 2, 183 -190. (2003)
12. Kuo-Chen Chou, Yu-Dong Cai:. Using GO-PseAA predictor to predict enzyme sub-class. Biochemical and Biophysical Research Communications, 325, 506–509. (2004)
13. Huang WL, Chen HM, Hwang SF, Ho SY.: Accurate prediction of enzyme subfamily class using an adaptive fuzzy k-nearest neighbor method. BioSystems 90, 405-413. (2007)
14. Hamilton O. Smith, Thomas M. Annau, Srinivasan Chandrasegran.: Finding sequence motifs in groups of functionally related proteins. Proc. Natl. Acad. Sci. USA 87, 826-830. (1990)
15. Lawrence S Hon, Ajay N Jain.: A deterministic motif finding algorithm with application to the human genome. Bioinformatics 22, 1047-1054. (2006)