Prediction of Enzyme Class by Using *Reactive Motifs* Generated from Binding and Catalytic Sites

Peera Liewlom, Thanawin Rakthanmanon, and Kitsana Waiyamai

Data Analysis and Knowledge Discovery Laboratory (DAKDL), Computer Engineering Department, Engineering Faculty, Kasetsart University, Bangkok, Thailand
{oprll, fengtwr, kitsana.w}@ku.ac.th

Abstract. The purpose of this research is to search for motifs directly at binding and catalytic sites called *reactive motifs*, and then to predict enzyme functions from the discovered reactive motifs. The main challenge is that the data of binding, or catalytic sites is only available in the range 3.34% of all enzymes, and many of each data provides only one sequence record. The other challenge is the complexity of motif combinations to predict enzyme functions.

In this paper, we introduce a unique process which combines statistics with bio-chemistry background to determine *reactive motifs*. It is consisting of *block scan filter*, *mutation control*, and *reactive site-group define* procedures. The purpose of *block scan filter* is to alter each 1-sequence record of binding or catalytic site, using similarity score, to produce quality blocks. These blocks are input to *mutation control*, where in each position of the sequences, amino acids are analyzed an extended to determine complete substitution group. Output of the *mutation control* step is a set of motifs for each 1-sequence record input. These motifs are then grouped using the *reactive site-group define* procedure to produce reactive motifs. Those reactive motifs together with known enzyme sequence dataset are used as the input to C4.5 learning algorithm, to obtain an enzyme prediction model. The accuracy of this model is checked against testing dataset. At 235 enzyme function class, the reactive motifs yield the best prediction result with C4.5 at 72.58%, better than PROSITE motifs.

Keywords: mutation control, sequence motif, reactive motif, enzyme function prediction, binding site, catalytic site, amino acid substitution group.

1 Introduction

An enzyme function or enzyme reaction mechanism is the combination of two main sub-functions: binding, and catalyzing. The parts in an enzyme sequence are called binding sites, and catalytic sites. A site is a short amino acid sequence. To perform one type of binding or catalyzing may be achieved by each of several short amino acid sequences. These sequences can be represented in one pattern (motif).

One of the most well-known collections of motif sequences is PROSITE [1]. PROSITE contains only 152 motifs of binding and catalytic sites, covering 396 out of 3,845 classes of enzyme functions. Therefore the insufficient of data is one of main

R. Alhajj et al. (Eds.): ADMA 2007, LNAI 4632, pp. 442–453, 2007.

challenges. In addition, one of the motifs can be a part of 46 enzyme functions, while 139 enzyme functions can have more than one of the motifs. These create complexity. Therefore, many methods [2,3,4,5] avoid the direct usage of motifs generated from binding and catalytic sites to predict enzyme functions. Those methods use other resources and need data in the form of blocks [6] or multiple sequence alignment [7], which contain very few sequences of binding and catalytic sites.

In this paper, we choose to develop the method to predict enzyme functions based on the direct usage of these binding and catalytic sites. Principal motivation is that information of enzyme reaction mechanism is very important for applied science, especially bioinformatics. We introduce a unique process to determine *reactive motifs* using *block scan filter*, *mutation control*, and *reactive site-group define*. The main step, *mutation control*, is a method based on motif patterns of PROSITE, which involve amino acid substitution, insertion-deletion, and conserved region, to generate amino-acid substitution group. For example, with the PROSITE motif [RK]-x(2,3)-[DE]-x(2,3)-Y, mutation in position 1 is a substitution [8,9] of amino acids R or K, while maintaining same function. The mutation in position 2 is a insertion-deletion (ins-dels/gap) of amino acids of 2 or 3 residues, and the last position is a conserved region Y necessary in most mutation sequences. In our work, only conserved region and substitution are used in the mutation control operation.

The amino acid substitution has been described in 2 paradigms; expert-based and statistic-based motifs. In the case of expert-based motifs such as PROSITE, the substitution is manually resulted from expert knowledge and bio-chemistry background. The main principle is that the different enzymes with the same reaction mechanism on binding sites and catalytic sites perform the same enzyme function [8]. Due to the need of expertise, motifs discovered by experts are in slow progress. In the case of statistic-based motifs such as EMOTIF [3], motifs are discovered using statistical methods. Therefore the fast predictions of enzyme functions can be achieved. Almost of the statistic-based motifs are not discovered directly from the binding sites or catalytic sites; but from the surrounding sites. Statistic-based motifs yield high enzyme function prediction accuracy. However, in certain applications, it is necessary to understand how motifs of these sites are combined to perform enzyme function. This is the reason why the statistic-based motifs cannot replace the expert-based motifs completely.

In this paper, we propose a method for searching motifs directly at binding and catalytic sites called *reactive motifs*. The proposed method combines statistics with bio-chemistry background in the similar way of expert working process. We develop a procedure called *block scan filter* to alter the 1-sequence record of binding or catalytic site to generate a block, which will be input to the *mutation control* step. As a result, 1-sequence record can produce one motif. The motifs generated from a set of input sequences are then grouped using the *reactive site-group define* procedure to produce reactive motifs. Those reactive motifs together with known enzyme sequence dataset are used as the input to C4.5 learning algorithm, to obtain an enzyme prediction model.

The following will be the details in reactive motifs discovery (phase I), and reactive motif-based prediction of enzyme class (phase II). The overall process is described in Fig. 1, and details are described in section 2 and 3. Experimental results, conclusion and discussion are given respectively in section 4 and section 5.

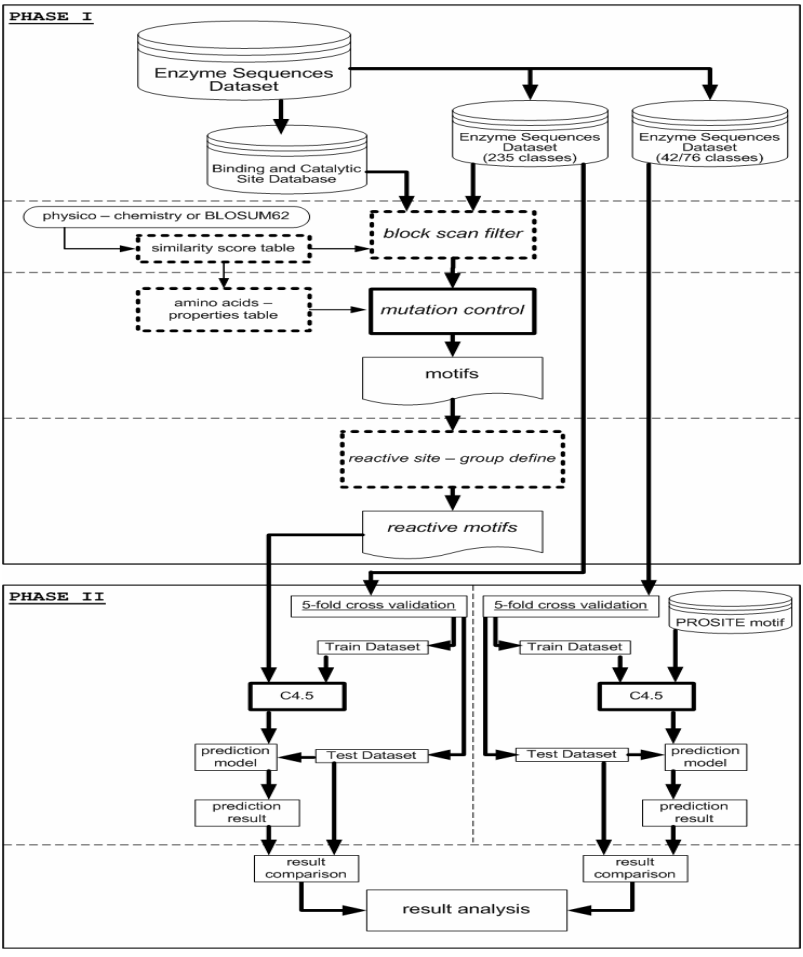


Fig. 1. Overview of enzyme function prediction using reactive motifs with mutation control

2 Reactive Motifs Discovery with Mutation Control (Phase I)

This phase consists of 4 steps; data preparation step, block scan filter step, mutation control step, and reactive site-group define step. The result is reactive motif representing each enzyme mechanism. These details are explained order in the next.

2.1 Data Preparation

We use the protein sequences data from the SWISSPROT part [10] in the UNIPROT database release 9.2, and the enzyme function class from ENZYME NOMENCLATURE [11] in the ENZYME NOMENCLATURE database release 37.0 of Swiss Institute of Bioinformatics (SIB). The enzyme protein sequences are grouped

to be used as a database, called *Enzyme Sequence Dataset*. Some of the enzyme proteins in the Enzyme Sequence Dataset provide the information of the amino acid position, which is a part of a binding or catalytic site. Setting the position as the center, a binding or catalytic sequence with the length of 15 amino acids, forming a binding or catalytic site, is retrieved from the enzyme protein sequence (See Fig. 2). These binding and catalytic sites are grouped and used as another database called *Binding and Catalytic Site Database*.

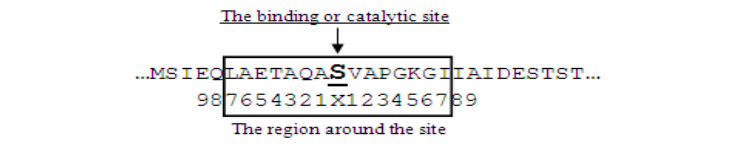


Fig. 2. Data preparation of the sequences around the binding and catalytic site

In this Binding and Catalytic Site Database, the sites are divided to subgroups. In case the sites are binding sites, the sites are in the same subgroup when they have the same reaction descriptions, which are the same substrate, the same binding method (i.e. via amide nitrogen), and the same type of amino acid(s). In case the sites are catalytic sites, the sites are in the same subgroup when they have the same mechanism (i.e. the same proton acceptor), and the same type of amino acid(s). There are in total 291 subgroups in this Binding and Catalytic Site Database. The sites in each subgroup will be used to scan to all related enzyme protein sequences in the *block scan filter* step in order to get quality blocks.

In a function class, if only one type of binding or catalytic site is found, the function class is also neglected. The reason is the enzyme classes having only one motif cannot represent the complexity of the sub-function combination. Only the function classes having enzyme members between 10 and 1000 are used. The rest of the classes are neglected. Therefore, the Enzyme Sequence Dataset covers 19,258 enzymes in 235 function classes. And the Binding and Catalytic Site Database covers 3,084 records of binding or catalytic sites with 291 enzyme reaction descriptions.

2.2 Block Scan Filter

Objective of the block scan filter step is to alter one record of binding or catalytic site data to form a block. This step is divided into two subtasks: the *similarity block scanning* and the *constraint filter*. The first subtask is to use only 1 record of binding or catalytic site to induce the related binding and catalytic sites in order to create a block. One record of binding site or catalytic site is used to scan over the related protein sequences, all protein sequences in enzyme functions that have the same site descriptions. Several similarity scores, such as BLOSUM62 [12], are given, while the record scans over. The part of the protein with the length of 15 amino acids, giving highest score will be stored in a block. The scanning is repeated to other related protein sequences. Therefore, the result is a block containing sets of highest score binding or catalytic sites.

From this block, some of the sites inside the block are filtered out using the second subtask, *constraint filter*. To achieve that, the sites in the block are sorted from highest scores to the lowest. Based on the works of Smith et.al. [13], a block has high quality when each site in the block having at least 3 positions presenting the same type of amino acids. This is the criteria to filter the block. Fig. 3. shows example of block members selection of the constraint filter subtask.

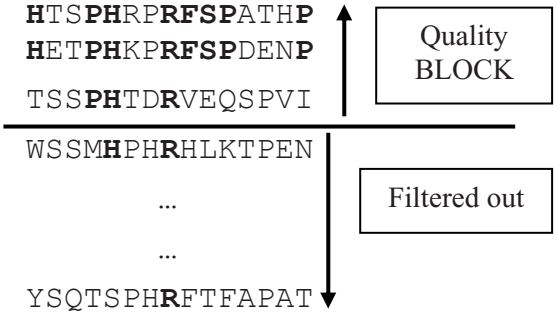


Fig. 3. Selection of members in the blocks using *constraint filter*

2.3 Mutation Control

An enzyme mechanism can be represented by several amino acid sequences of binding or catalytic sites. Therefore specific positions in sequences necessary for controlling the properties of the enzyme mechanism shall have common or similar properties. In some positions, they are of the same types of amino acids, which are called *conserved regions*. In some positions, there are many types of amino acids, however having similar properties. All amino acids in the same position are grouped with respect to the mutation in biological evolution and the resulting group is called *substitution group*. The characteristics of the substitution group can be of two types submitted from patterns of PROSITE motif:

(1) The substitution group shall have some common properties representing by [], for example [ACSG].

(2) The amino acids having prohibited properties shall not be included in the group. This prohibition is represented by {}, for example {P}, meaning any amino acids but P.

We call *mutation control* when the substitution group at each position of binding or catalytic sites is controlled by the above characteristics. Thus, mutation control regarding to biological evolution is important for enzyme mechanism to function. The objective of mutation control is to determine complete substitution group at each position in sequences. The mutation control is formalized using concept lattice theory is explained in [16]. In the following, we give an example to illustrate mutation control process.

Table 1. The physico-chemistry table representing the background knowledge of amino acids properties

	Small	Tiny	Proline	Polar	Charge	Positive	Negative	Hydrophobic	Aromatic	Aliphatic
A	X	X						X		
C	X	X		X				X		
D	X			X	X		X			
E				X	X		X			
F								X	X	
G	X	X						X		
H				X	X	X		X	X	
I								X		X
K				X	X	X		X		
L								X		X
M								X		
N	X			X						
P	X		X							
Q				X						
R				X	X	X				
S	X	X		X						
T	X			X				X		
V	X							X		X
W				X				X	X	
Y				X				X	X	

For the first pattern [ACSG], their common properties using background knowledge from physico-chemistry (see Table 1) is {small, tiny} which are necessary for enzyme mechanism to function. For the second pattern {P}, the prohibited property is {proline}, for which other amino acids do not have. The prohibited property blocks the enzyme mechanism to function. We call *boundary properties*, the complement of the prohibited property that do not block the enzyme mechanism.

For example (see Fig. 3), at position 1 of BLOCKS, the original substitution group is {H,T}. Their common properties are hydrophobic and polar. It follows that the amino acid group having the common properties is {H, T, W, Y, K, C}. All properties representing {H,T} are the boundary properties, or the properties - polar charge positive hydrophobic aromatic and aliphatic. Any amino acid having the other properties may be the prohibited properties which blocks the enzyme mechanism. Thus, the amino acids having the boundary properties are {H, T, F, K, M, N, Q, R, W, Y}. The complete substitution group controlled by the common properties and the boundary properties can be obtained by intersecting the amino acids generated from the common properties and the boundary properties. From this example, we obtain the complete substitution group as {H, T, W, Y, K}. This process is repeated with all other positions of the quality block. The result is a reactive motif from one binding or catalytic site.

However, the output motifs from using *block scan filter* and *mutation control* should be generated using the same background knowledge. In case of using physico-chemistry table in the *mutation control* step, similarity score table transformed from the physico-chemistry table should be used in the *block scan filter* step. Similarly, when using BLOSUM62 table as similarity score table in *block scan filter step*, we should use amino acids properties table transformed from BLOSUM62 table at the *mutation control* step.

The similarity score table transformed from Physico-Chemistry is given in Table 2. The score is given in relation to the number of the same properties, for example, if two amino acids have 3 same properties, the similarity score is 3. For example, amino acids A and C have properties {small, tiny, hydrophobic} and {small, tiny, polar, hydrophobic}, the similarity score is the shared properties weight by $1 = |\{\text{small, tiny, hydrophobic}\} \cap \{\text{small, tiny, hydrophobic}\}| = 3$. However in case of pairing the same amino acid type, the score is weighed more than one, in our case, it is weighed by 4.

The reason is to give higher score for conserved region. For example, the similarity score of amino acid A and itself is 4 x |{small, tiny, hydrophobic} ∩ {small, tiny, hydrophobic}| = 12.

Table 2. The similarity score table transformed from physico-chemistry table of amino acids properties

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	12	3	1	0	1	3	1	1	1	1	1	1	1	0	0	2	2	2	1	1
C	3	16	2	1	1	3	2	1	2	1	1	2	1	1	1	3	3	2	2	2
D	1	2	16	3	0	1	2	0	2	0	0	2	1	1	2	2	2	1	1	1
E	0	1	3	12	0	0	2	0	2	0	0	1	0	1	2	1	1	0	1	1
F	1	1	0	0	8	1	2	1	1	1	1	0	0	0	0	0	1	1	2	2
G	3	3	1	0	1	12	1	1	1	1	1	1	1	0	0	2	2	2	1	1
H	1	2	2	2	2	1	20	1	4	1	1	1	0	1	3	1	2	1	3	3
I	1	1	0	0	1	1	1	8	1	2	1	0	0	0	0	0	1	2	1	1
K	1	2	2	2	1	1	4	1	16	1	1	1	0	1	3	1	2	1	2	2
L	1	1	0	0	1	1	1	2	1	8	1	0	0	0	0	0	1	2	1	1
M	1	1	0	0	1	1	1	1	1	1	4	0	0	0	0	0	1	1	1	1
N	1	2	2	1	0	1	1	0	1	0	0	8	1	1	1	2	2	1	1	1
P	1	1	1	0	0	1	0	0	0	0	0	1	8	0	0	1	1	1	0	0
Q	0	1	1	1	0	0	1	0	1	0	0	1	0	4	1	1	1	0	1	1
R	0	1	2	2	0	0	3	0	3	0	0	1	0	1	12	1	1	0	1	1
S	2	3	2	1	0	2	1	0	1	0	0	2	1	1	1	12	2	1	1	1
T	2	3	2	1	1	2	2	1	2	1	1	2	1	1	1	2	12	2	2	2
V	2	2	1	0	1	2	1	2	1	2	1	1	1	0	0	1	2	12	0	0
W	1	2	1	1	2	1	3	1	2	1	1	1	0	1	1	1	2	0	12	3
Y	1	2	1	1	2	1	3	1	2	1	1	1	0	1	1	1	2	0	3	12

On the other hand, the amino acids properties table transformed from BLOSUM62 table consist of 3 steps. Firstly, BLOSUM62 is transformed to binary table using threshold at zero. The score more than or equal to zero is replaced by 1, and the score less than zero is set to zero. Then, a property is set by the greatest group of amino acids sharing binary 1. Last step, all the new properties are put a table in relation to the amino acids. All steps are described in Fig. 4.

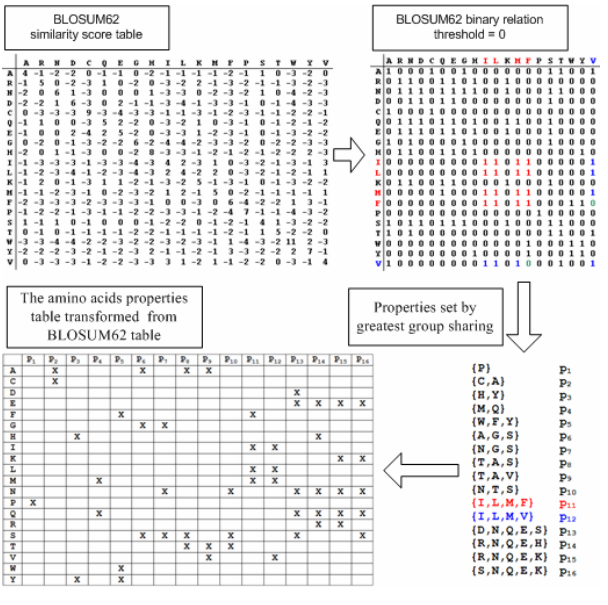


Fig. 4. The amino acids properties table transformed from BLOSUM62 table

These two background knowledge types give different potentials. The background knowledge based on BLOSUM62, in general, is better statistically, while the background knowledge based on physico-chemistry yields motifs closer to PROSITE.

From Fig. 3, we can discover two different reactive motifs from 1 binding or catalytic site with respect to the different background knowledge. Using physico-chemistry table, we obtain [HTWYK] [CDENQST] [CNST] P H [KNQRT] [DNP] R [FILMV] [DENQS] . [ACDGNST] . . . as output motif. Using BLOSUM62, we obtain . . [ST] P H . . R . [ENS] . . . . . as output motif.

2.4 Reactive Site – Group Define

From the previous step, motifs are produced from different records of the same binding or catalytic function, by definition, are redundant, and should be grouped together and represent as one motif, namely *reactive motif*. It means that the 291 subgroups in the Binding and Catalytic Site database will yield 291 reactive motifs.

Although motifs are retrieved from the same original binding or catalytic sites in the same subgroup of the Binding and Catalytic Site Database; they can have different binding structures to the same substrate. In other words, there are many ways to "fit and function". Therefore these motifs, in some cases, can be rearranged to several reactive motifs. The separate method called *conserved region group define* is based on the conserved region, where the motifs with the same amino acids at the same positions of conserved regions are grouped together. As the results, 1,328 reactive motifs are achieved by BLOSUM62 tool, and 1,390 by physico-chemistry table tool. This grouping process is called the *reactive site – group define*.

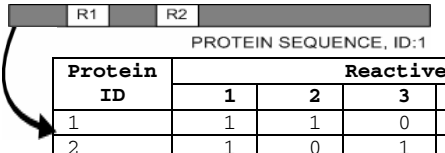
3 Reactive Motif-Based Prediction of Enzyme Class (Phase II)

In this phase, the problem is to construct an enzyme prediction model using reactive motifs together with known Enzyme Sequence Dataset (train data set). The efficiency of our prediction model is compared with the one of PROSITE, the original pattern subscribed by mutation control to create reactive motif automatically.

Concerning data preparation for phase II, the motifs and enzyme classes in PROSITE are selected for the comparison purpose. In PROSITE, there are 152 motifs of binding and catalytic sites. To be comparable, the same conditions used with reactive motif are applied. However using the condition, which the function classes having members between 10 and 100, covers very small number of motifs (36 motifs in 42 functions, 2,579 sequences) and yields very low accuracy. Therefore we use the function classes having members between 5 and 1000 instead, which covering 65 motifs in 76 Enzyme Classes (2,815 sequences).

To construct the prediction model, given a set of the motifs (reactive motifs or PROSITE motifs), we aim to induce classifiers that associate the motifs to enzyme functions. As suggested in [14], any protein chain can be mapped into a representation based attributes. Such a representation supports efficient function of data-driven algorithms, which represent instances as classified part of fixed set of attributes. In our case, an enzyme sequence is represented as a set of reactive motifs (or PROSITE motif for comparison purpose).

Suppose that from phase I, *N* reactive motifs have been obtained. Each sequence is encoded as an *N*-bit binary pattern where the i^{th} bit is 1 if the corresponding reactive motif is present in the sequence; otherwise the corresponding bit is 0. Each *N*-bit, sequence is associated with an EC number (Enzyme Commission Number). A training set is simply a collection of *N*-bit binary patterns each of which has associated with it, an EC number. This training set can be used to train a classifier which can then be used to assign novel sequences to one of the several EC-numbers represented in the training set. The reactive motif-based representation procedure is given in Fidg. 5.



Protein ID	Reactive motif						Class (EC)
	1	2	3	4	….	N	
1	1	1	0	0	….	0	3.2.1.1
2	1	0	1	0	….	0	3.2.1.1
3	0	0	0	1	….	1	4.1.2.13
….	….	….	….	….	….	….	….
M	0	1	0	0	….	1	2.6.1.1

Fig. 5. Reactive Motif-Based Representation of Enzymes

In this paper, we use Weka [15], the machine learning suit, to compare the efficiency of different enzyme function prediction models. C4.5 decision tree (J4.8 Weka's implementation) has been used as a prediction learner in order to assess efficiency of the reactive motifs used for predicting enzyme functions.

4 Experimental Results

In this part, we present the results of the efficiency and the quality of the reactive motifs to predict enzyme functions. The results are divided to 2 sections 1) the prediction accuracy comparison between reactive motifs resulted from different background knowledge (BLOSUM62 or physico-chemistry table), 2) The quality of each reactive motif.

4.1 Prediction Accuracy Comparison Between Reactive Motifs Resulted from Different Background Knowledge

This section compares the accuracy of prediction between different enzyme function prediction models, which resulted from the reactive motifs, which are generated from different background knowledge. The reactive motif generated from BLOSUM62 is called *BLOSUM – reactive motif*. The reactive motif generated from Taylor's physico-chemistry table is called *physicochemistry – reactive motif*. The reactive motifs with out substitution group element are used as reference of reactive motif, that retrieve from conserved region of reactive motif generated from BLOSUM62, called *conserved amino acid – reactive motif*. In addition, the prediction accuracy of enzyme function prediction model from PROSITE motifs is presented. The dataset we used

covers 235 enzyme function classes, 19,258 protein sequences. The enzyme function prediction models are created by learning algorithm C4.5 with 5 fold – cross validation. The results are presented in the table 3 and 4.

In case the *conserve region -group define* step is not applied, the prediction model with using BLOSUM – reactive motifs gives the best result: 68.69% accuracy. The prediction model with using physicochemistry – reactive motifs with application of conserve region-group define gives the best result: 72.58% accuracy, however, the accuracies of all models are very close.

Table 3. The maximum scale comparison among the enzyme function prediction systems (19,258 sequences, 235 functions)

Reactive site - group define	Reactive motif					
	Conserved amino acid		BLOSUM		Physicochemistry	
	# motif	C4.5 (%)	# motif	C4.5 (%)	# motif	C4.5 (%)
From Binding and Catalytic Site Database	291	60.84	291	**68.69**	291	64.38
Conserved region - group define	1324	70.57	1328	71.66	1390	**72.58**

Table 4. The maximum scale of the enzyme function prediction system with PROSITE motifs

Selected function class with condition of #Members	# Functions	# Motifs	# Sequences	C4.5 (%)
Between 10 and 1000	42	36	2579	37.15
Between 5 and 1000	76	65	2815	**67.25**

The accuracy of the prediction model retrieved from PROSITE motifs gives the best result of 67.25%.

4.2 Quality of Discovered Reactive Motifs

In case the learning algorithm is not used, the quality efficiency of motifs/reactive motifs to represent the sub-functions of binding or catalytic sites are measured and compared. The quality is represented by 2 values: *coverage value*, and *motifs found per enzyme sequence*. The coverage value is the percentage of the motifs that relevant to enzyme sequences in all related enzyme classes. From the sequences, which motifs/reactive motifs cover, each sequence is checked on how many motifs are matched, and the average value from all sequences is calculated, called motifs found per enzyme sequence.

From table 5, the higher the coverage value is the better. However, the motifs found per sequence, theoretically, should close to 2, because one enzyme at least has one type of binding site and one type of catalytic site. The reactive motifs using physico-chemistry background knowledge gives the result closest to PROSITE, both coverage value and motifs found per sequence.

Table 5. Show the quality values of the reactive motifs and the PROSITE motif

Motif type	# Class	# seq	# seq not match any motif	Coverage value (%)	Motif found per enzyme sequence
PROSITE	42	2579	1752	32.07	1.5562
PROSITE	76	2815	590	79.04	1.6431
Conserved amino acid - Reactive motif	235	19258	59	99.69	27.8416
BLOSUM - Reactive motif	235	19258	665	96.55	6.6293
Physicochemistry - Reactive motif	235	19258	2772	85.61	3.4724

5 Conclusion and Discussion

The process introduced here yields good results (~70% accuracy of enzyme function prediction), and can solve the main problems such as the insufficient of data: binding sites and catalytic sites (~5.8% in our dataset). The reactive motifs using physico-chemistry background knowledge gives the best results, although the coverage value is not satisfied, the reactive-motifs found per enzyme sequence is very good. It means the motifs are very specific. The improvement of accuracy caused from conserved region group define shows that the details in the mechanism descriptions are not complete. The quality of the descriptions of binding and catalytic sites should be improved.

The proposed reactive motif discovery process can be applied using other types of background knowledge. Using other background knowledge such as HMM profile to classify protein domain or family in another interesting future work.

References

1. Bairoch, A.: PROSITE: a dictionary of sites and patterns in proteins. Nucleic Acids Res. 19, 2241–2245 (1991)
2. Sander, C., Schneider, R.: Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins Struct. Funct. Genet. 9, 56–68 (1991)
3. Huang, J.Y., Brutlag, D.L.: The EMOTIF database. Nucleic Acids Res. 29, 202–204 (2001)
4. Eidhammer, I., Jonassen, I., Taylor, W.R.: Protein structure comparison and structure patterns. Journal of Computational Biology 7(5), 685–716 (2000)
5. Bennett, S.P., Lu, L., Brutlag, D.L.: 3MATRIX and 3MOTIF: a protein structure visualization system for conserved sequence. Nucleic Acids Res. 31, 3328–3332 (2003)
6. Henikoff, S., Henikoff, J.G.: Automated assembly of protein blocks for database searching. Nucleic Acids Res. 19, 6565–6572 (1991)
7. Barton, G.J.: Protein multiple sequence alignment and flexible pattern matching. Methods Enzymol (183), 403–428 (1990)
8. Taylor, W.R.: The classification of amino acid conservation. J. Theor. Biol. 119(2), 205–218 (1986)

9. Wu, T.D., Brutlag, D.L.: Discovering Empirically Conserved Amino Acid Substitution Groups in Databases of Protein Families. In: Proc. Int. Conf. Intell. Syst. Mol. Biol., vol. (4), pp. 230–240 (1996)
10. Bairoch, A., Apweiler, R.: The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. 28, 45–48 (2000)
11. Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB): Enzyme Nomenclature. Recommendations 1992. Academic Press (1992)
12. Henikoff, S., Henikoff, J.G.: Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA (89), 10915–10919 (1992)
13. Smith, H.O., Annau, T.M., Chandrasegaran, S.: Finding sequence motifs in groups of functionally related proteins. Proc. Natl. Acad. Sci. 87(2), 826–830 (1990)
14. Diplaris, S., Tsoumakas, G., Mitkas, P.A., Vlahavas, I.: Protein Classification with Multiple Algorithms. In: Bozanis, P., Houstis, E.N. (eds.) PCI 2005. LNCS, vol. 3746, Springer, Heidelberg (2005)
15. Frank, E., Hall, M., Trigg, L., Holmes, G., Witten, I.H.: Data mining in bioinformatics using Weka. Bioinformatics 20(15), 2479–2481 (2004)
16. Liewlom, P., Rakthanmanon, M.P., Waiyamai, K.: Concept Lattice-based Mutation Control for Reactive Motif Discovery. DAKDL technical report, Faculty of Engineering, Kasetsart University, Thailand