Concept Lattice–Based Mutation Control for Reactive Motifs Discovery

Kitsana Waiyamai, Peera Liewlom, Thanapat Kangkachit, and Thanawin Rakthanmanon

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

Abstract. We propose a method for automatically discovering *reactive motifs*, which are motifs discovered from binding and catalytic sites, which incorporate information at binding and catalytic sites with bio-chemical knowledge. We introduce the concept of *mutation control* that uses amino acid substitution groups and conserved regions to generate complete amino acid substitution groups. Mutation control operations are described and formalized using a concept lattice representation. We show that a concept lattice is efficient for both representations of bio-chemical knowledge and computational support for mutation control operations. Experiments using a C4.5 learning algorithm with reactive motifs as features predict enzyme function with 72% accuracy compared with 67% accuracy using expert-constructed motifs. This suggests that automatically generating reactive motifs are a viable alternative to the time-consuming process of expert-based motifs for enzyme function prediction.

Keywords: mutation control, concept lattice, sequence motif, reactive motif, enzyme function prediction, binding site, catalytic site.

1 Introduction

There are many statistic-based motif methods for enzyme function prediction capable of high accuracy; however, most of these methods [2,3,4,5] avoid the direct usage of motifs generated from binding and catalytic sites to predict enzyme function prediction. These methods use other resources from surrounding sites that contain very few sequences of binding and catalytic sites. In certain applications, it is necessary to understand how motifs of binding and catalytic sites are combined in order to perform enzyme function prediction. This is a reason why the statistic-based motifs cannot completely replace expert-identified motifs. In this paper, we develop a method to predict enzyme functions based on direct usage of binding and catalytic sites. Motifs discovered from binding and catalytic sites are called *reactive motifs*. The principal motivation is that different enzymes with the same reaction mechanism at binding and catalytic sites frequently perform the same enzyme function. In previous work [16], we introduced a unique process to discover reactive motifs using *block scan filtering*, *Mutation Control*, and *Reactive Site-Group Definition*. The main step in reactive

T. Washio et al. (Eds.): PAKDD 2008, LNAI 5012, pp. 767–776, 2008.
© Springer-Verlag Berlin Heidelberg 2008

motif discovery is mutation control whose objective is to determine a complete substitution group for each position in the sequences, such that the substitution group contains all possible amino acids that can be substituted.

In this paper, we show that the concept lattice provides an efficient representation of various types of bio-chemical background knowledge and efficient computational support for the operations of mutation control. We propose a method to construct an amino-acid property context from background knowledge which is Taylor Physico-Chemistry table [8]. From the amino-acid property context, the concept lattice representing a hierarchy of amino-acid substitution groups sharing the same properties is constructed. Each concept represents a substitution group; lattice operators are applied to obtain complete substitution groups. Reactive motifs generated from the concept-lattice mutation control step are used as input to the C4.5 learning algorithm to obtain the enzyme prediction model. The reactive motifs and PROSITE [1] motifs separately are used as training data for the C4.5 learning model, which is then evaluated using a test dataset containing 19,258 amino acid sequences of 235 known enzyme functions. The learning algorithm using reactive motifs as training data accurately identified 72.6% of the test sequences, compared to 67.25 % accuracy for PROSITE.

The overall process of reactive motif discovery is described in section 2. Section 3 gives details of the concept lattice-based mutation control; experimental results are presented in section 4, and conclusions are given in section 5.

2 Reactive Motifs Discovery with Mutation Control

In this section, we present an overall process of reactive motif discovery, consisting of three steps: *data preparation and block scan filtering*, *mutation control*, and *reactive site-group definition*. More details of reactive motif discovery process can be found in [16].

2.1 Data Preparation and Block Scan Filtering

In the data preparation step, we use an *enzyme sequence dataset* [10,11] that covers 19,258 enzyme sequences of 235 functions. Within this enzyme sequence dataset, we use sequences containing binding or catalytic sites. Designating the binding or catalytic site position as the center, binding or catalytic site sub-sequences are retrieved, each of length 15 amino acids, as shown in Fig. 1. These binding and catalytic site sub-sequences form a *binding and catalytic site* database. Sub-sequences in the binding and catalytic site database are then clustered into subgroups based on their reaction descriptions. There are in total 291 subgroups.

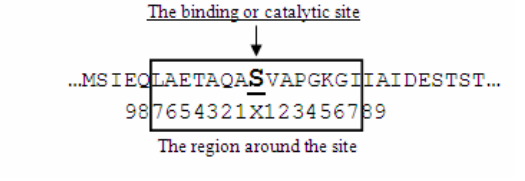

Fig. 1. Sequence with length of 15 amino acids around the binding and catalytic site

The purpose of the *block scan filtering* step is to alter each record of the binding and catalytic site database. For each binding or catalytic site sub-sequence, the dataset is scanned for all other sequences having the same site description, and a sequence similarity score is computed using amino-acid similarity scores such as BLOSUM62 [12]. The sequences are ranked according to similarity scores; then a block member filtering method [13] is applied. A block is designated as high quality when each site in the block has at least 3 positions presenting the same type of amino acids, as shown in Fig. 2.

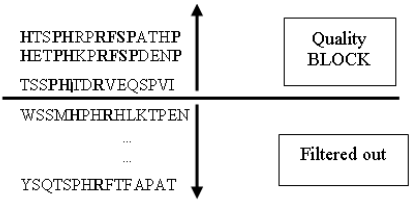

Fig. 2. Block member filtering to obtain a high quality block

2.2 Mutation Control

An enzyme mechanism can be represented by several binding or catalytic site subsequences. Therefore specific positions in sequences that control the properties of the enzyme mechanism have common or similar properties. Some positions in all sequences contain the same type of amino acids; these positions are called *conserved regions*. Other positions may have many types of amino acids, but 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 a *substitution group*. Therefore, a substitution group is a set of amino acids having common or similar properties that can be substituted at a specific position in a block. There are two kinds of substitution groups, represented by patterns as in the PROSITE motifs:

(1) A group of amino acids having some common properties; the substitutable amino acids are listed in brackets, for example [HT].
(2) Amino acids having *prohibited properties* cannot be included at a position in the group. Prohibited amino acids are listed in braces, for example {P}, meaning any amino acid except P.

Mutation control constructs a motif consisting of the complete substitution group or conserved region from each position in the sequence. Using the results of the block scan filter step, all amino acids in the same position are compared and analyzed. Mutation control extends each amino acids substitution group to include all amino acids having common characteristics, identified using the Taylor physico-chemistry table (Table 1), to create a *complete substitution group*. This extension process is described next.

Table 1. Physico-chemistry table representing 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	

A complete substitution group is constructed by examining both the *common properties* and *boundary properties* at a given position. In some positions, there may be many types of amino acids that yield the same enzyme reaction mechanism. These amino acids have common or similar properties. For example, the amino-acids substitution group [HT] has *Polar* and *Hydrophobic* as common properties, which are necessary for an enzyme mechanism to function.

The *prohibited properties* are all the properties that are not found by any member of the substitution group. For example, the prohibited properties of [HT] are *Tiny*, *Negative,* and *Aliphatic*. The *boundary properties* set is the complement of the prohibited properties. The boundary properties and common properties are used together to identify the complete substitution group.

To be certain that a given substitution group contains all possible amino-acids that can be substituted, the mutation control extends each substitution group to include all amino acids that have *all* the common properties and *only* properties in the boundary set (i.e. no prohibited properties). For example, complete substitution group for [HT] is [HTWYK]. This is the greatest amino acid substitution group that has all common properties and the only properties they have are boundary properties. This complete substitution group is determined at all other positions of the quality block to produce a motif. For the quality block in Fig. 2, we obtain the motif [HTWYK] [CDENQST] [CNST] P H [KNQRT] [DNP] R [FILMV] [DENQS] [ACDGNST] . . .

The source of background information can be used in block scan filtering and mutation control should be the same. For example, if the BLOSUM62 table is used as the similarity score table in *block scan filtering step*, the amino acids properties table transformed from BLOSUM62 should be used in the *mutation control* step. More details about background knowledge transformation can be found in [16].

2.3 Reactive Site – Group Definition

From the previous step, motifs produced from different records of the same binding or catalytic functions are, by definition, redundant. They are grouped together and represented as one *reactive motif* in a grouping process called *reactive site–group definition*. 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". As a result of this step, 1,328 reactive motifs are constructed using the BLOSUM62 data and 1,390 using the Taylor physico-chemistry table.

3 Concept Lattice–Based Mutation Control for Complete Substitution Group Discovery

In this section, we apply concept lattice theory [17,18] to mutation control in order to determine complete substitution groups. From the amino acids context, the concept lattice is generated, where concepts are constructed as amino acids substitution groups sharing common properties. The generated concept lattice represents hierarchy of amino acids substitution groups sharing common properties. From this lattice, mutation control operations are performed to determine complete amino-acid substitution groups. We start by giving some basic definitions of concept lattices as applied to mutation control. Then, concept lattice-based mutation control operations are defined.

3.1 Basic Definitions

Amino acid properties context: *An amino acid properties context is a triple ($\sum$, P, R), where $\sum$ and P are finite sets of amino acids and properties, and $R \subseteq \sum x P$ is a binary relation. eRp denotes that the amino acid $e \in \sum$ is in relation R to the property $p \in P$, if e has the property p (or e verifies property p).*

Concept: *A concept is a pair (Extent, Intent) where Extent $\subseteq$ E, Intent $\subseteq$ P and f(Extent)=Intent and g(Intent) =Extent. Let L be a set of all concepts formed from the context ($\sum$, P, R), and let $c \in L$. Hence, c is formed by two parts: an extent representing a subset of $\sum$ (here, amino acids), denoted as Extent(c), and an intent representing the common properties between this subset of amino acids, denoted as Intent(c).* For example, ({A,C,G},{small, tiny, hydrophobic}) is a concept of the context in Table 1. This means that there are no more than three amino acids possessing at least all properties in {small, tiny, hydrophobic} and sharing at most these properties in common. The concept's extent is an amino-acid substitution group sharing similar properties.

Amino acid properties concept lattice: *An amino acid properties concept lattice is a concept lattice L = (L,$\leq$) of an amino acid properties context ($\sum$, P , R) , is a complete lattice of concepts derived from the amino acid properties context. The lattice structure imposes:*

- *a partial ordering on concepts such that for concepts c1, c2 $\in$ L, c1 $\leq$ c2, iff Extent(c1) $\subseteq$ Extent(c2) or, equivalently, Intent(c2) $\subseteq$ Intent(c1).*
- *any concept subset of L has one greatest subconcept (the Meet element) and one least superconcept (the Join element).*

Theorem. *Let ($\sum$,P,R) be a context, let L be a concept lattice of concepts derived from ($\sum$,P,R) and S $\subseteq$ L. The Meet(S) and Join(S) elements are given as follows:*

$$Meet(S) = (\bigcap_{c \in S} Extent(c), f(g(\bigcup_{c \in S} Intent(c)) \quad (1)$$

$$Join(S) = (g(f(\bigcup_{c \in S} Extent(c)), \bigcap_{c \in S} Intent(c)) \quad (2)$$

3.2 Complete Amino-Acids Substitution Group Discovery

In this section, we present a method for finding complete amino acid substitution group at a given position of a block of amino acids resulted from the block scan filtering step (section 2.1). Our method works in 4 steps. First, it starts by finding smallest object concept for each amino acid in the amino acid-properties lattice. Then, it uses those concepts to find candidate substitution groups having the greatest common properties and having the greatest boundary properties. Finally, it returns the common amino acids of both substitution groups as the complete amino-acid substitution group.

3.2.1 Finding Amino Acid Concepts

Each amino acid in the same position of a block is used for finding its introduction concept in the amino acid-properties lattice called *amino-acid concept* [19]. Considering Fig. 3, ({H}, {aro,hyd,cha,pol,pos}) and ({T,C} ,{hyd,sma,pol}) are introduction concepts of amino-acids H and T.

Fig. 3. Shows two candidate-substitution groups of amino acids {H, T} which are represented by gray nodes. The black node represents candidate substitution group having the greatest common properties, derived from the gray edges, while candidate substitution group having the boundary properties, represented by dotted node, can be derived from the dotted edges.

3.2.2 Finding Candidate Substitution Group Having Common Properties

According to an important characteristic of a substitution group (described in section 2.2), complete substitution group should have common properties. In order to determine the substitution group having common properties at most or *greatest set of common properties*, the lattice operator *Join(S)* is applied where S is a set of amino-acid concepts derived from the previous section. *Join(S)* returns a concept whose intent contains greatest common properties of S and whose extent is a candidate substitution group.

In the following, we show how the greatest common properties of amino acids {H,T} and its candidate substitution group can be determined. From the previous step, we obtained ({H},{aro,hyd,cha,pol,pos}) and ({T,C} ,{hyd,sma,pol}) as amino-acid concepts represented as gray nodes in the Fig. 3. Then, we use them as input to the *Join(S)* operator. ({W,H,Y,K,T,C},{hyd,pol}) is the result of *Join(S)* whose extent represents candidate substitution group of amino acids {H,T}.

3.2.3 Finding Candidate Substitution Group Having Boundary Properties

According to the definition of a substitution group (described in Section 2.2), a complete substitution group should exclude any amino acid having the prohibited properties that prevent the enzyme mechanism function. The substitution group having the greatest set of boundary properties is the result of the union of the extent of all super-concept of the lattice operator *Meet(S)*, where S is a set of amino-acid concepts as described in Section 3.2.1. In the case that *Meet(S)* produces a concept whose intent contains any prohibited properties, a virtual boundary concept will be used instead. The intent of the virtual boundary concept includes only the greatest boundary properties and its extent is an empty set. A virtual boundary concept can be formally defined as follows:

Definition: *Let* $(\sum, P, R)$ *be a context, L be a concept lattice derived from* $(\sum, P, R)$, *and* $S \subseteq L$. *A concept* $(\varnothing, \bigcup_{c \in S} Intent(c))$ *is called a virtual boundary concept if Meet* $(S) = (\varnothing, I)$ *and* $I \not\subset \bigcup_{c \in S} Intent(c)$.

In the following, we show how the greatest set of boundary properties of amino acids {H,T} and their candidate substitution group can be determined. From Section 3.2.1, we obtain S = {({H},{aro,hyd,cha,pol,pos}), ({T,C} ,{hyd,sma,pol})} as a set of amino-acid concepts represented by gray nodes in the Fig. 3. Then, *Meet(S)* results the bottom concept ({},{sma,tin,aro,neg,ali,hyd,cha,pol,pos}). In this case, the intent of result concept contains prohibit properties such as {tin, pros, neg, ali}. Thus, a virtual boundary concept ({},{aro,hyd,cha,pol,pos,sma}) is created. We then link it as the immediate predecessor concept of the bottom concept. Then, we determine its immediate predecessor concepts by choosing the immediate predecessor concepts of the bottom concept having no prohibited properties, which is the set of concepts {({H},{aro,hyd,cha,pol,pos}), ({T,C},{hyd,cha,pol,pos})}, represented by a dashed node in Fig. 3. Finally, from the set of super-concepts of the virtual boundary concept, we select only object concepts. Then, the union of the extent of those object concepts is the substitution group having boundary properties {H,T,W,Y,F}.

3.2.4 Complete Amino Acid Substitution Group

Once both candidate substitution groups are extracted from the previous step, a complete amino acid substitution group can be determined by finding the common amino acids appearing in both substitution groups. From Fig. 3, amino acids having common properties are {W,H,Y,K,T,C}; while amino acids having the boundary properties are {H,T,W,Y,F}. Thus, the amino acids that appear in both substitution groups form the complete substitution group {H,T,W,Y} of amino acids {H, T}, as required.

4 Experimental Results

We performed experiments using a dataset containing 19,258 protein sequences that covers 235 enzyme functions, using the C4.5 learning algorithm with a 5-fold cross validation.

The accuracy of the enzyme function prediction models is shown in Table 2. Each prediction model is constructed using reactive motifs generated from different background knowledge. The model constructed with reactive motifs generated using BLOSUM62 is called *BLOSUM – reactive motif*. The model constructed with reactive motifs generated using Taylor's physico-chemistry table is called *physico-chemistry – reactive motif*. The reference model, called *conserved amino acid – reactive motif*, is constructed using reactive motifs without a substitution group. These reactive motifs are generated from conserved regions using BLOSUM62. In case the *conserved region-group definition* step is not applied, the *BLOSUM – reactive motifs* model gives the best results with 68.69% accuracy. The prediction model using physico-chemistry – reactive motifs with application of conserved region-group definition gives the best accuracy, 72.58%; however, the accuracies of all models are very close.

Table 3 shows the prediction accuracy of enzyme function prediction model, with respect to different class members using PROSITE motifs. The accuracy of the prediction model retrieved from PROSITE motifs gives the best accuracy of 67.25%.

Table 2. Accuracy comparison among function prediction models using reactive motifs

Reactive site– group definition	Reactive motifs					
	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 definition	1324	70.57	1328	71.66	1390	**72.58**

Table 3. Accuracy of function prediction models using PROSITE motifs

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

5 Conclusions and Discussion

In this paper, we show that concept lattice is an efficient representation of biochemistry background knowledge and an efficient computational support for mutation control operations. To obtain an enzyme prediction model, reactive motifs generated from the concept lattice based mutation control step are used as the input to C4.5 learning algorithm. Our enzyme prediction model yields good results (~70% accuracy of enzyme function prediction) and can overcome problems such as lack of protein or enzyme functional information; only about ~5.8% in our dataset contain information about binding and catalytic sites. The reactive motifs using physico-chemistry background knowledge give the best results; although the coverage value is not satisfied, the number of reactive motifs found per enzyme sequence is very good. It indicates the motifs are very specific.

The limited improvement in accuracy observed when using the conserved region group definition indicates that the details of the mechanism descriptions need further improvement. Improving the quality of the descriptions of binding and catalytic sites would, in the authors' view, further increase the accuracy of enzyme function prediction using reactive motifs.

Acknowledgement. Thanks to J. E. Brucker for his reading and comments of this paper.

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. Proc. Int. Conf. Intell. Syst. Mol. Biol. (4), 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. U S A 87(2), 826–830 (1990)
14. Diplaris, S., Tsoumakas, G., Mitkas, P.A., Vlahavas, I.: Protein Classification with Multiple Algorithms. In: Proc. of 10th Panhellenic Conference in Informatics, Volos, Greece, November 21-23. LNCS. Springer, Heidelberg (2005)
15. Frank, E., Hall, M., Trigg, L., Holmes, G., Witten, I.H.: Data mining in bioinformatics using Weka. Bioinformatics 2004 20(15), 2479–2481 (2004)
16. Liewlom, P., Rakthanmanon, P., Waiyamai, K.: Prediction of Enzyme Class using Reactive Motifs generated from Binding and Catalytic Sites. In: Alhajj, R., Gao, H., Li, X., Li, J., Zaïane, O.R. (eds.) ADMA 2007. LNCS (LNAI), vol. 4632. Springer, Heidelberg (2007)
17. Wille, R.: Restructuring lattice theory: an approach based on hierarchies of concepts. In: Rival, I. (ed.) Ordered sets, Dordrecht–Boston, pp. 445–470 (1982)
18. Waiyamai, K., Taouil, R., Lakhal, L.: Towards an object database approach for managing concept lattices. In: Embley, D.W. (ed.) ER 1997. LNCS, vol. 1331, pp. 299–312. Springer, Heidelberg (1997)
19. Arévalo, G., Berry, A., Huchard, M., Perrot, G., Sigayret, A.: Performances of Galois Subhierarchy-building Algorithms. In: Kuznetsov, S.O., Schmidt, S. (eds.) ICFCA 2007. LNCS (LNAI), vol. 4390, pp. 166–180. Springer, Heidelberg (2007)