FUZZY CONCEPT LATTICE-BASED APPROACH FOR REACTIVE MOTIFS DISCOVERY

Thanapat Kangkachit and Kitsana Waiyamai
Department of Computer Engineering, Faculty of Engineering, Kasetsart University, Bangkok, Thailand

Keywords: Complete substitution group, Fuzzy concept lattice, Reactive motifs, Enzyme function classification, Binding and catalytic site, Amino acid sustitution matrix, Biochemical knowledge.

Abstract: Reactive motifs are short conserved regions discovered from binding and catalytic sites of enzymes sequences. Thus, reactive motifs provide more biological meaning than statistic-based motifs because they are directly extracted from where the chemical reaction mechanism occurs. Main problem of discovering reactive motifs is that only 4.94% enzymes sequences contain sites information. To overcome this problem, we present fuzzy concept lattice-based (FCL-based) method for discovering more general reactive motifs by incorporating biochemical knowledge. Fuzzy concept lattices are used to represent both binary and multi-value biochemical knowledge. The fuzzy concept lattice Join operator is applied to determine complete substitution groups that obtains more general reactive motifs. Experiments are conducted among different methods of determining complete substitution groups: FCL-based, concecpt lattice-based (CL-based) and similarity-based method. Experimental results show that FCL-based method significantly outperforms other methods in term of coverage value and F-measure with SVM learning algorithm. Therefore, fuzzy concept lattice provides more efficient computational support for complete substitution groups operation than that of other existing methods.

1 INTRODUCTION

Reactive motifs are short conserved regions discovered from binding and catalytic sites of enzymes sequences. Compared with statistic-based motifs (Sander and Schneider, 1991; Eidhammer et al., 1999; Huang and Brutlag, 2001; Bennett et al., 2003), enzyme function classification model using reactive motifs gives the better accuracy with explanation in terms of reactive motifs combination. Compared with expert-based motifs i.e. PROSITE (Bairoch, 1993), the performance of classification model using reactive motifs is more efficient in term of accuracy due to a small number of occurrences in protein sequences of those expert-based motifs. Main problem in discovering reactive motifs is the lack of binding and catalytic sites information, only 4.94% of enzymes sequences contain binding and catalytic sites information in the UNiProtKB/Swiss-Prot Version 9.2 (Bairoch and Apweiler, 2000). (Waiyamai et al., 2008) have proposed a concept lattice-based reactive motifs discovery method called CL-based. They introduced the concept of mutation control determine a complete amino acid substitution group for each position in the sequences, such that the substitution group contains all possible amino acids that can be substituted. Concept lattice operators have been defined to support mutation control operations. The proposed technique yields good results (70 % accuracy of enzyme function classification) and can overcome problems such as lack of information about binding and catalytic sites. However, only binary-value context is supported, conversion method is needed to support both binary and multi-value knowledge. Moreover, the occurrences of reactive motifs need to be improved to obtain better accuracy of enzyme function classification model.

This paper presents FCL-based approach for reactive motifs discovery. Main objective is to further increase the occurrences of reactive motifs in enzymes sequences dataset and to improve their quality by incorporating various types of biochemical knowledge. Both binary and multi-value biochemical knowledge is formally constructed in a unique structure fuzzy concept lattice that overcomes the problem of information loss while converting multi-value knowledge into binary-value context. Fuzzy concept lattices are then used to determine complete substitution groups in reactive motifs. As result, more general reactive motifs are generated. We also show that FCL-based approach provides efficient computational support for complete substitution groups determination in reac-

Kangkachit T. and Waiyamai K..
FUZZY CONCEPT LATTICE-BASED APPROACH FOR REACTIVE MOTIFS DISCOVERY.
DOI: 10.5220/0003787503260330
In *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms* (BIOINFORMATICS-2012), pages 326-330
ISBN: 978-989-8425-90-4
Copyright © 2012 SCITEPRESS (Science and Technology Publications, Lda.)

tive motifs discovery. The quality of FCL-based reactive motifs can be expressed through the high performance of enzyme function classification model using them as features to learning algorithm i.e. C4.5 (Quinlan, 1993) and SVM (Boser et al., 1992; Cristianini and Shawe-Taylor, 2010).

The experiments are conducted among different complete substitution groups determination methods: FCL-based, CL-based and similarity-based methods. The performance of reactive motifs is measured using the coverage value (occurrences of motifs in enzyme sequences dataset) and F-measure with SVM and C4.5. The experimental results show that FCL-based method gives better performance than other methods in terms of coverage value, F-measure, Precision and Recall. In the following, we describe the process of reactive motifs discovery using fuzzy-concept lattice. Experimental results and conclusion are presented in section 3 and section 4 consequently.

2 FUZZY CONCEPT LATTICE-BASED APPROACH FOR REACTIVE MOTIFS DISCOVERY

In this section, the overall process of reactive motifs discovery is presented. It is composed of three steps: *data preparation and block scan filtering, complete substitution groups determination, and reactive site group definition* as the same framework as introduced in (Waiyamai et al., 2008).

2.1 Data Preparation and Block Scan Filtering

In this step, we collect enzyme sequences dataset (Bairoch, 1993; Apweiler et al., 2004), covers 22,637 sequences of 237 functions, while only 4.94% of enzyme sequences contain binding or catalytic sites. By considering sites position as center, enzyme subsequences having 15 amino acids length; *initial sites patterns* are extracted. This is based on the 21-atoms average substrate size in BRENDA database (Schomburg et al., 2004) which corresponds to 7 amino acids (3 atoms per 1 amino acid) on each side of the sites position.

To solve the problem of lack information at binding and catalytic sites, the generalization is performed at each position of the *initial sites patterns* using a block scan filtering. Then, a scan operation is performed to retrieve all the sub-sequences having the same site description to generate *sequences blocks*.

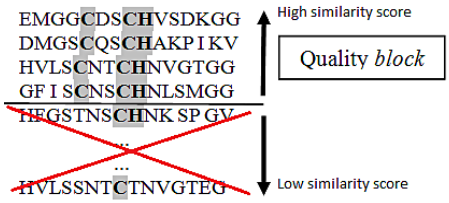


Figure 1: Enzyme sub-sequence filtering to obtain high quality *block*.

Figure 2: A *block pattern* obtained from the MSA operation on a quality block.

Sequences are ranked according to amino-acid similarity scores using BLOSUM62 table (Henikoff and Henikoff, 1992). In Figure 1, to obtain high quality block, each sub-sequence in the block must has at least 3 positions presenting the same type of amino acids as proposed in (Smith et al., 1990) for the high quality motifs.

To discard some positions that may not involve in protein functional mechanism, a multiple sequences alignment (MSA (Ramu et al., 2003)) is performed to all sub-sequences in each high quality *block*. Finally, a *block pattern* is extracted as a representative block in the form of regular expression e.g. **x[FMV]x[GS]C[DQN][ST]CHxxxxx**, where *x* be any amino acid, [ ] be a set of possible amino acids at a given position, *called substitution group*, as shown in Figure 2.

Due to the lack sites information, there will possibly be other amino acids in some substitution group positions of a *block pattern*. Thus, determining complete substitution groups to generate more general and high quality reactive motifs from a set of *block patterns* is required.

2.2 Complete Substitution Groups Determination

In the following, we explain how the complete substitution groups are determined using FCL-based method. First, both binary and multi-value biochemical knowledge from various sources can be formally constructed in a unique fuzzy concept lattice structure (Yahia and Jaoua, 2001; S., 2003; Elloumi et al., 2004). The values in each properties of biochemi-

cal knowledge are normalized to [0,1] and then formulated as an amino acid properties context to construct fuzzy concept lattice. Due to the fact that biochemical knowledge can contain a large number of amino acid properties, FCL-based uses (Nakai et al., 1988) approach to select proper properties. A hierarchical cluster analysis is used to investigate the relationships among amino acid properties (circles with index) and represent them using a minimum spanning tree as shown in Figure 3(a). Figure 3(b) shows the selection of properties in largest cluster (region with dots) to be input to the amino acid properties context. Once constructed from the amino acid properties context, the fuzzy concept lattice structure is able to express a hierarchy of concepts which represent amino-acid substitution groups sharing the common properties. Then, complete substitution groups operations can be defined based on the fuzzy concept lattice Join operator. More details of fuzzy concept lattice can be found in (Yahia and Jaoua, 2001; S., 2003; Elloumi et al., 2004).

Here, we show how FCL-based method works to determine complete substitution groups. In Figure 2, the substitution group [FMV] at 2^{nd} position of a block pattern is used as an example. According to amino acid properties context in Table 3(b), the common properties of this group can be determined by the minimum value of all properties i.e. $(\#383^{0.52}, \#96^{0.60}, \#159^{0.15}, \ldots, \#27^{0})$. Then, a set of amino acids **{F,M,V,I,L}** is extracted as a complete substitution group with respect to their common properties of this substitution group. Finally, we obtain pattern **x[FMVIL]x[GS]C[DQN][ST]CHxxxxx** as a set of complete reactive motifs.

2.3 Reactive Sites Groups Generation

Enzymes can have different catalytic and binding structures to fit and function the same substrate. Thus, complete reactive motifs can be grouped together according to their sites description as specified in UNiProtKB/Swiss-Prot database. As result, 291 *reactive sites groups* which correspond to the reactive motifs groups are generated and used as input features to build the enzyme functions classification model.

3 EXPERIMENTAL RESULTS

In the following, the results of different types of reactive motifs: FCL-based, baseline (without use of background knowledge), CL-based, similarity-based, are compared using a dataset containing 22,637 protein sequences with 237 enzyme functions collected from UNiProtKB/Swiss-Prot Version 9.2. The different types of reactive motifs are explained in Section 3.1. The performance comparison is conducted in terms of the reactive motif generalization and the prediction model accuracy.

3.1 Different Types of Reactive Motifs

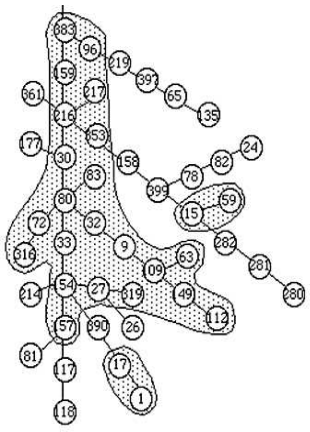
Here, implementation details of the different types of reactive motifs are given.

- Baseline method discovers reactive motifs without use of biochemical knowledge. The propose of this method is to compare how efficient the other methods gain in using proper background knowledge.
- CL-based method integrates binary-value biochemical knowledge via a concept lattice, and performs mutation control operations to determine complete substitution groups to generate reactive motifs. More details of CL-based can be found in (Waiyamai et al., 2008).
- CL-based* method extends CL-based method by performing a multiple sequence alignment operation on quality blocks (refer to step 2.1) before determining complete substitution groups.
- Similarity-based method uses amino acid substitution matrices instead of biochemical properties knowledge. For each substitution group, the minimum similarity score obtained by any pairs of amino acids in such group is computed. The other amino acids are determined as complete substitution group if the similarity scores between them and any amino acid in the substitution group are higher than or equal to minimum similarity score. For example, the minimum similarity score of substitution group **[MFV]** is -1 derived from similarity score between F and V using BLOSUM62. Then, we can extract [**MFV**HILWYARCQKST] as a complete substitution group.

3.2 Evaluation of Reactive Motifs Discovery Methods

The performance of reactive motifs is evaluated in terms of their occurrences in enzymes sequences dataset through *coverage value* and the quality of prediction model using them as input features measured by using *F-Measure, Precision and Recall*. Table1 compares performance of the different methods for reactive motifs discovery.

In term of coverage value, FCL-Based and similarity-based methods produce significantly higher value compared to no use of knowledge. It means that



(a)

	#383	#96	#159	#217	#216	#353	#30	#83	#80	#72	#316	#33	#32	#9	#109	#63	#149	#112	#154	#157	#319	#27
A	0.89	0.28	0.03	-7.02	-9.48	86.60	0	2.04	1.00	89.09	0.13	115.00	91.50	52.60	88.30	2.50	0.28	31.00	0.77	0.77	93.70	0
R	0.88	0.10	0.20	-10.13	-16.23	162.20	0	6.24	6.13	174.20	-5.00	225.00	202.00	109.10	181.20	7.50	0.34	124.00	3.72	2.38	250.40	1
N	0.89	0.25	0.10	-9.42	-12.48	103.30	1	4.37	2.95	132.12	-3.04	160.00	135.20	75.70	125.10	5.00	0.31	56.00	1.98	1.45	146.30	1
D	0.87	0.21	0.10	-9.30	-12.14	97.80	1	3.78	2.78	133.10	-2.23	150.00	124.50	68.40	110.80	2.50	0.33	54.00	1.99	1.43	142.60	1
C	0.85	0.28	0.10	-8.19	-12.21	132.30	0	3.41	2.43	121.15	-2.52	135.00	117.70	68.30	112.40	3.00	0.11	55.00	1.38	1.22	135.20	0
Q	0.82	0.35	0.10	-10.04	-13.69	119.20	0	3.53	3.95	146.15	-3.84	180.00	161.10	89.70	148.70	6.00	0.39	85.00	2.58	1.75	177.70	1
E	0.84	0.33	0.10	-10.47	-13.82	113.90	1	3.31	3.78	147.13	-3.43	190.00	155.10	84.70	140.50	5.00	0.37	83.00	2.63	1.77	182.90	1
G	0.92	0.17	0.03	-5.46	-7.59	62.90	1	1.00	0.00	75.07	1.45	75.00	66.40	36.30	60.00	0.50	0.28	3.00	0.00	0.58	52.60	0
H	0.83	0.21	0.10	-12.15	-17.55	155.80	0	5.66	4.66	155.16	-5.61	195.00	167.30	91.90	152.60	6.00	0.23	96.00	2.76	1.78	188.10	1
I	0.76	0.82	0.19	-9.51	-15.61	158.00	0	3.49	4.00	131.17	-2.77	175.00	168.80	102.00	168.50	5.50	0.12	111.00	1.83	1.56	182.20	1
L	0.73	1.00	0.19	-10.52	-15.73	164.10	0	4.45	4.00	131.17	-2.64	170.00	167.90	102.00	168.50	5.50	0.16	111.00	2.08	1.54	173.70	2
K	0.97	0.09	0.20	-9.67	-12.37	115.50	0	4.87	4.77	146.19	-3.97	200.00	171.30	105.10	175.60	7.00	0.59	119.00	2.94	2.08	215.20	1
M	0.74	0.74	0.19	-10.42	-15.70	172.90	0	4.80	4.43	149.21	-3.83	185.00	170.80	97.70	162.20	6.00	0.08	105.00	2.34	1.80	197.60	1
F	0.52	2.18	0.39	-12.49	-20.50	194.10	0	6.02	5.89	165.19	-3.74	210.00	203.40	113.90	189.00	6.50	0.10	132.00	2.97	1.90	228.60	1
P	0.82	0.39	0.17	-8.65	-11.89	92.90	0	4.31	2.72	115.13	0.00	145.00	129.30	73.60	122.20	5.50	0.46	32.50	1.42	1.25	0.00	0
S	0.96	0.12	0.03	-7.78	-10.52	85.60	0	2.70	1.60	105.09	-1.66	115.00	99.10	54.90	88.70	3.00	0.27	32.00	1.28	1.08	109.50	0
T	0.92	0.21	0.10	-8.76	-12.37	106.50	0	3.17	2.60	119.12	-2.31	140.00	122.10	71.20	118.20	5.00	0.26	61.00	1.43	1.24	142.10	0
W	0.20	5.70	0.56	-14.42	-26.17	224.60	0	5.90	8.08	204.24	-8.21	255.00	237.60	135.40	227.00	7.00	0.15	170.00	3.58	2.21	271.60	1
Y	0.49	1.26	0.39	-12.36	-20.23	177.70	0	6.72	6.47	181.19	-5.97	230.00	203.60	116.20	193.00	7.00	0.25	136.00	3.36	2.13	239.90	1
V	0.85	0.60	0.15	-8.78	-13.87	141.00	0	3.17	3.00	117.15	-2.05	155.00	141.70	85.10	141.40	5.00	0.22	84.00	1.49	1.29	157.20	0

(b)

Figure 3: AAIndex: physicochemical properties **(a)** Representation of 3 clusters (region with dots) using minimum spanning tree **(b)** Amino acid-properties context derived from the largest cluster in (a).

Table 1: Coverage value, F-measure, recall and precision comparison among different types of reactive motifs.

Reactive motifs types	Coverage(%)	F-Measure[4] SVM	F-Measure[4] C4.5	Recall[2] SVM	Recall[2] C4.5	Precision[3] SVM	Precision[3] C4.5
Similarity-based method							
- BLOSUM62	94.55	0.699	0.662	**0.709**	0.672	**0.745**	0.675
- BLOSUM80	93.80	0.691	0.666	0.700	0.675	0.742	0.686
- PAM30	**95.43**	**0.705**	0.661	0.683	**0.677**	0.720	**0.687**
- PAM70	94.33	0.689	0.654	0.698	0.664	0.735	0.668
- PAM250	94.35	0.699	**0.667**	0.708	0.676	0.743	0.684
Mutation control-based methods							
- FCL-Based[1]	**98.02**	**0.746**	0.635	**0.751**	0.646	**0.766**	0.646
- CL-Based*	91.81	0.672	**0.660**	0.684	0.676	0.722	0.691
- CL-Based	64.84	0.590	0.586	0.549	0.545	0.615	0.609
Baseline method	91.91	0.662	0.660	0.675	0.670	0.713	0.684

[1] Using integrated biochemical knowledge [2] Recall = TP / (TP + FN) [3]Precision = TP / (TP + FP)
[4] F-Measure = 2*Precision*Recall / (Precision + Recall);

both methods efficiently utilize biochemical knowledge in order to produce more general reactive motifs. Moreover, FCL-based method provide best computational support for generating general reactive motifs due to its highest coverage value.

In term of prediction model accuracy, FCL-based method gives highest precision, recall and F-measure with SVM. It means that FCL-based method also produces highest true positive(TP) value but lowest false positive(FP) and false negative(FN) values. In contrast, the other methods aim to produce high precision but low recall that affected F-Measure i.e. high FP and FN values. However, there is no significantly different of all measures obtained by FCL-based method and other methods with C4.5.

In summary, with highest coverage, F-Measure, precision and recall values, FCL-based method provides more general reactive motifs while retains accuracy of the prediction model by reducing the effect of FP and FN values.

4 CONCLUSIONS AND DISCUSSION

Main problem of discovering reactive motifs is that only 4.94% enzymes sequences contain sites information. To overcome this problem, we present fuzzy concept lattice-based (FCL-based) method for discovering more general reactive motifs by incorporating biochemical knowledge. Fuzzy concept lattices are used for both representing binary and multi-value biochemical knowledge, and determining complete sub-

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

stitution groups that produces more general reactive motifs. Used as input features of SVM, generated FCL-based reactive motifs provide highest coverage, F-measure, precision and recall values without affecting the FP and FN values in the prediction model.

Rule-based learning methods can be investigated to provide meaningful and understandable information to biologists. Among rule-based methods, associative classification technique (Liu et al., 1998; Li et al., 2001; Belohlvek et al., 2007) is recognized to be more accurate over traditional classification techniques i.e. C4.5 for very large number of classes to predicted. In the future work, we will increase both accuracy and explanatory ability of the protein function classification model using reactive motifs as input features to an associative classification.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the Thailand Research Fund (TRF) and Kasetsart University for financial support through the Royal Golden Jubilee Ph.D. scholarship program (1.0.KU/49/A.1).

REFERENCES

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N., and Yeh, L.-S. L. (2004). Uniprot: the universal protein knowledgebase. *Nucleic Acids Research*, 32(Database-Issue):115–119.

Bairoch, A. (1993). The prosite dictionary of sites and patterns in proteins, its current status. *Nucleic Acids Research*, 21(13):3097–3103.

Bairoch, A. and Apweiler, R. (2000). The swiss-prot protein sequence database and its supplement trembl in 2000. *Nucleic Acids Res*, 27:49–54.

Belohlvek, R., Baets, B. D., Outrata, J., and Vychodil, V. (2007). Inducing decision trees via concept lattices. In *CLA*, volume 331 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Bennett, S. P., Lu, L., and Brutlag, D. L. (2003). 3matrix and 3motif: a protein structure visualization system for conserved sequence motifs. *Nucleic Acids Research*, 31(13):3328–3332.

Boser, B. E., Guyon, I., and Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In *COLT*, pages 144–152.

Cristianini, N. and Shawe-Taylor, J. (2010). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press.

Eidhammer, I., Jonassen, I., and Taylor, W. R. (1999). Structure comparison and structure patterns. *JOURNAL OF COMPUTATIONAL BIOLOGY*, 7:685–716.

Elloumi, S., Youssef, C. B., and Yahia, S. B. (2004). The fuzzy classifier by concept localization in a lattice of concepts. In *Proceedings of the CLA 2004 International Workshop on Concept Lattices and their Applications (CLA)*.

Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919.

Huang, J. Y. and Brutlag, D. L. (2001). The emotif database. *Nucleic Acids Research*, 29(1):202–204.

Li, W., Han, J., and Pei, J. (2001). Cmar: Accurate and efficient classification based on multiple class-association rules. In *Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM)*, pages 369–376.

Liu, B., Hsu, W., and Ma, Y. (1998). Integrating classification and association rule mining. In *Knowledge Discovery and Data Mining*, pages 80–86.

Nakai, K., Kidera, A., and Kanehisa, M. (1988). Cluster analysis of amino acid indices for prediction of protein structure and function. *Protein Eng*, 2(2):93–100.

Quinlan, J. R. (1993). *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Ramu, C., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G., and Thompson, J. D. (2003). Multiple sequence alignment with the clustal series of programs. *Nucleic Acids Research*, 31(13):3497–3500.

S., K. (2003). Cluster based efficient generation of fuzzy concepts. In *Neural Network World*, pages 521–530.

Sander, C. and Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins: Structure, Function, and Genetics*, 9(1):56–68.

Schomburg, I., Chang, A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G., and Schomburg, D. (2004). BRENDA, the enzyme database: updates and major new developments. *Nucleic acids research*, 32(Database issue):D431–433.

Smith, O., T., A. M., and S., C. (1990). Finding sequence motifs in groups of functionally related proteins. *Proceedings of the National Academy of Sciences*, 87(2):826–830.

Waiyamai, K., Liewlom, P., Kangkachit, T., and Rakthanmanon, T. (2008). Concept lattice-based mutation control for reactive motifs discovery. In *PAKDD*, pages 767–776.

Yahia, S. B. and Jaoua, A. (2001). *Discovering knowledge from fuzzy concept lattice*, pages 167–190. Physica-Verlag GmbH, Heidelberg, Germany, Germany.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS