Genomics & Informatics Vol. 1(2) 75-79, December 2003

Interpretation of Association Networks among Protein Sequence Motifs

Hye J. Kam, Junehawk Lee, Doheon Lee* and Kwang H. Lee

Bio-Information System Laboratory in Department of BioSystems, Korea Advanced Institute of Science and Technology, Daejeon, Korea.

Abstract

Every protein can be characterized by either a distinct motif or a combination of motifs. Nevertheless, little is known about the relationships among (more than two) the motifs. Some of the proteins in the world are share motifs for evolutional or other biological benefits - they can save energy, time and resource for controlling and managing a variety of proteins. In some cases of motifs, the tendency is quite common and they can act the 'hub' motif of a network of the motif associations. The hubs are structurally and functionally important in themselves and also important in disease-related mutations. They will be highly resistant mutation to conserve their functions. But, in case of the a rare mutation, mutations on the position of hub can more easily cause fatal diseases.

Keywords: Association, Association networks, Motif, Rule mining

Introduction

To understand the biological activities of a cell, it is essential to characterize the functions of the proteins in that specific cell. From a particular function of a protein, there can be complex biological processes. Among many approaches for protein characterization, there is a method using the proteins' basic building units - amino acids.

As the size of protein sequence database grows, it becomes more powerful to predict proteins' functions by sequence comparison and there have been those kinds of works. In this paper, we used the protein sequence motifs that are representative features of a protein and their functions rather than comparing whole protein sequences (we are going to use the term 'motif' as a concept that includes 'functional domain'). Every protein can be characterized by either a distinct motif or a combination of motifs. And, to understand motif association is crucial for understanding the nature and the extent of biochemical functions (Deng *et al.*, 2002).

Moreover, motif combinations are interesting because combining with other motifs, a motif can show more complicated and specified function to a protein.

Nevertheless, little is known about the relationships among (more than two) the motifs. There are several motif finding algorithms and protein motif databases such as PROSITE (Flaquet *et al.*, 2002), PRINTS (Attwood *et al.*, 2003), PFAM (Bateman *et al.*, 2002) and SMART (Letunic *et al.*, 2002) using those algorithms. In this paper, we use d the InterPro database (Mulder *et al.*, 2003) that provides integrated and non-redundant protein-to-motif data and found association rules on the motifs using a data mining algorithm.

Some of the protein in the world share motifs for evolutional or other biological benefits - they can save energy, time and resource for controlling and managing a variety of proteins. In some of the motifs, the tendency of sharing is quite common and they can act as the 'hub' motif for a network of motif associations. The hubs are structurally and functionally important in themselves and are also important in disease-related mutations. They are highly resistant to conserve their functions. But, in the case of a rare mutation on the position of hub can more easily cause fatal diseases.

In this paper, we obtained association rules among the protein sequence motifs, and the found hub motifs from the networks that were construct from the rules. After that, we examined structures of the motif-containing proteins and searched the disease-related databases.

Association rule mining

Association rule mining finds interesting association or correlation relationships among a large set of data items. With massive amounts of data continuously being collected and stored, many industries are becoming interested in mining association rules from their databases. A typical example of association rule mining is market basket analysis.

* Corresponding author:
E-mail dhlee@bioif.kaist.ac.kr, Tel +82-42-869-4316, Fax +82-42-869-8680
Accepted 3 December 2003

Basic concepts

Let $J = \{i_1, i_2, \cdots i_m\}$ be a set of items (Han and Kamber, 2001). Let D, the task-relevant data, be a set of database transaction where each transaction T is a set of items such that $T \subseteq J$. Each transaction is associated with an identifier, called TID. Let A be a set of items. A transaction T is said to contain A if and only if $A \subseteq T$. An association rule is an implication of the form $A \Rightarrow B$, where $A \subset J$, $B \subset J$, and $A \cap B = \emptyset$. The rule $A \Rightarrow B$ holds in the transaction set D with support s, where s is the percentage of transactions in D that contain $A \cup B$ (i.e., both A and B). This is taken to be the probability, $P(A \cup B)$. The rule $A \Rightarrow B$ has confidence c in the transaction set D if c is the percentage of transactions in D containing A that also contain B. This is taken to be the conditional probability, $P(B|A)$. That is,

$$Support(A \Rightarrow B) = P(A \cup B)$$
$$Confidence(A \Rightarrow B) = P(B|A)$$

Association rule mined using a support-confidence framework are useful for many applications. However, the support-confidence framework can be misleading in that it may identify a rule $A \Rightarrow B$ as interesting when, in fact, the occurrence of A does not imply the occurrence of B. The occurrence of itemset A is independent of the occurrence of itemset B if $P(A \cup B) = P(A)P(B)$; otherwise itemset A and B are dependent and correlated as events. This definition can easily be extended to more than two itemsets. The correlation ('lift' in this paper) between the occurrence of A and B can be measured by computing,

$$Lift = Corr\, A,B = P(A \cup B) / (P(A)*P(B))$$

Association rule study in bioinformatics

Recently, there have been some approaches in biological fields with association rule mining. The subjects are protein-protein interaction, micro-array data analysis, literature-based gene identification and intra-protein motifs like ours (Horng *et al.*, 2002). The intra-protein motif study was actually a correlation study. The association rules did not shown in the paper and there wasn' t interesting analysis about the causal correlations, neither.

Differences from classification and correlation mining

Knowledge discovery through the association rule mining method is quite different from the classification method. Though the rules that are generated from a decision tree can be look like the general association rules, there are more than two statistical values only for association rules. As we mentioned before, association rules are considered interesting if they satisfy both a minimum support and a minimum confidence threshold. And more over, the values make it possible to obtain the non-deterministic 'tendency' of rules that can be opportunities for finding new information.

Mining association is also different from the correlation study. Association rules are built based on pre-measured correlation and given statistical meaning with those values like support and confidence. Association rules can include all of the information in correlations with the 'lift' value and can reproduce the correlations of themselves. And, the association rules can provide new information about the relationships of motifs.

Material & Method

Materials

We used the Swiss-Prot protein sequence database (v41.0) and the InterPro database (v6.0). The InterPro database has the integrated and non-redundant protein-to-motif data (Table 1).

Methods

For extracting the associations among the protein motifs, we used IBM Intelligent Miner for Data. The protein-to-motif data were built from the InterPro database with IDs to DB2. And, we set the minimum threshold values both a minimum support (0.06%) and a minimum confidence (50%).

An example

Here is an association rule simplified form of the real association rules of the motif data; $(m_1 \wedge m_2 \wedge \cdots \wedge m_k) \rightarrow m_{k+1}$. Symbol mi means a protein motif, and the arrow shows a subordinate relationships in math. Therefore, the symbolized association rule, $(m_1 \wedge m_2 \wedge \cdots \wedge m_k) \rightarrow m_{k+1}$, has meaning; "if one or more motif(s) m_1, m_2, $\cdots$, m_k are in a protein, then the motif m_{k+1}would exist in protein."

Results

Result of association rule mining

There were 141 networks that satisfied both a minimum support (0.06%) and a minimum confidence (50%) threshold. The size of network varied from group to group and the largest one had 38 motifs, the smallest one had only 2 motifs. From them we selected the largest one (with 38 motif set) as example of these networks. More than 2,400 rules were consisted that network and we picked a sub-network by choosing a herb motif randomly out for more sophisticate study. The sub-network was composed of 15 motifs and the hub of network was [IPR003593]: AAA

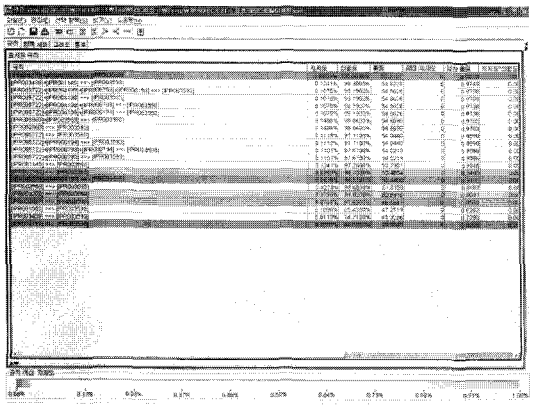

Fig. 1. Association rules of a selected sub-network.

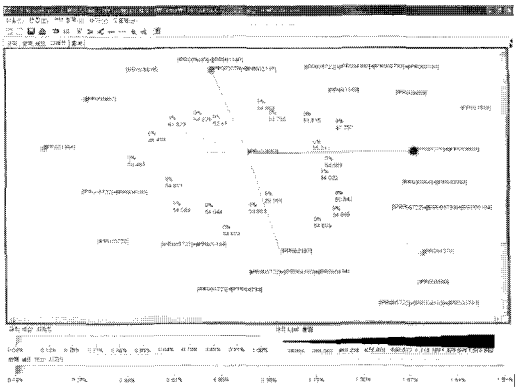

Fig. 2. Graph of sub-network that has [IPR003593] as hub.

Table 1. InterPro release 6.0 (March 2003).

Database	Version	Entries
Swiss-Prot	41.0	122564
PRINTS0	35.0	1750
TrEMBL	23.0	830525
PFAM	8.0	5193
PROSITE pattern	17.37	1605
PROSITE profile	N/A	150
ProDom	2002	1021
InterPro	6.0	7751
SMART	3.4	654
TIGRFAMs	2.1	1614

Table 2. InterPro motifs in sub-network and their annotations.

InterPro ID	Annotation
194	H+-transporting two-sector ATPase, alpha/beta subunit, central region
793	H+-transporting two-sector ATPase, alpha/beta subunit, C-terminal
897	GTP-binding signal recognition particle (SRP54) G-domain
1140	ABC transporter, transmembrane region
1270	Chaperonin clpA/B
1553	RecA bacterial DNA recombination protein
1984	Peptidase family S16
2078	Sigma-54 factor interaction domain
2197	Helix-turn-helix, Fis-type
3439	ABC transporter
3593	AAA ATPase
3959	AAA ATPase, central region
3960	AAA-protein subdomain
4100	H+-transporting two-sector ATPase, alpha/beta subunit, N-terminal
5722	ATP synthase F1, beta subunit

ATPase. The number of underlining association rules in the network were 21 (Fig. 1, 2).

Motif annotation table (from InterPro)

InterPro is an integrated documentation resource for protein families, domains and sites. InterPro combines a number of databases (referred to as member databases) that use different methodologies and a varying degree of biological information on the well-characterized proteins to derive the protein signatures. By uniting the member databases, InterPro capitalizes on their individual strengths, producing a powerful integrated diagnostic tool (Table 2).

Discussion

Newly identified relations:

[IPR001270], [IPR003960], [IPR001984], [IPR003439], [IPR002197], [IPR001140] with [IPR00353439], [IPR002197], [IPR001140] with [IPR003593]

In InterPro motif database, there were relations among sereral motifs: PARENT/CHI LD and CONTAINS/FOUND IN. The 'parent/child' relationship is used to indicate true protein family/subfamily relationships. In all cases a protein sequence match to the child entry implies a match to the parent and signatures for the parent and child entries must overlap. And, the 'contains/found in' relationship is used to indicate domain composition.

Some domains can be found in more than one type of protein or family of proteins, but is not a subtype in the family sense.

We identified four overlapped motifs: [IPR001270], [IPR003960], [IPR0019 84], [IPR003439] with [IPR0035 93]. And they represented similar function, 'ATP-binding' which is a kind of nucleotide. Moreover two more functional links were identified: [IPR002197], [IPR001140] with [IPR003593]. [IPR002197] stands for helix-turn-helix, which

binds to DNA. And [IPR001140] is ABC transporter transmembrane region motif.

Structural view

[IPR003593] is an integrated motif from the SMART database to InterPro. It indicates [SM00382] in the SMART motif database. From the structural information in SMART, we choose a structure (PDB id: 1BMF) that is obtained from an ATPase of bovine. We mapped [IPR003593] ([SM000382]) into that structure using NCBI Cn3D tool and got the site of that hub motif (Fig.3).

Mutation information

SMART database also provides OMIM curated human diseases associated with missense mutations within the AAA domain (IPR003593). There were 12 human proteins and 28 OMIM entries that are related to this hub motif (Table 3).

Table 3. 28 OMIM entries that are related to this hub motif.

drenoleukodystrophy protein (ALDP)	
OMIM 300100	Adrenoleukodystrophy ; Adrenomyeloneuropathy
Peroxisome assembly factor-2 (PAF-2) (Peroxisomal-type ATPase 1) (Peroxin-6).	
OMIM 601498	Peroxisomal biogenesis disorder, complementation group 4
Peroxisome biogenesis factor 1 (Peroxin-1) (Peroxisome biogenesis disorder protein 1)	
OMIM 602136	Zellweger syndrome-1
OMIM 214100	Adrenoleukodystrophy, neonatal
OMIM 202370	Refsum disease, infantile
OMIM 266510	-
Retinal-specific ATP-binding cassette transporter (RIM ABC transporter) (RIM protein) (RMP) (Stargardt disease protein)	
OMIM 601691	Stargardt disease-1
OMIM 248300	Retinitis pigmentosa-19
OMIM 601718	Cone-rod dystrophy 3 ; Macular dystrophy, age-related, 2
OMIM 153800	Fundus flavimaculatus
OMIM 248200	-
DNA repair protein RAD51 homolog 1 (hRAD51) (HsRAD51).	
OMIM 179617	{Breast cancer, susceptibility to}
OMIM 114480	-
Sulfonylurea receptor 1	
OMIM 600509	Persistent hyperinsulinemic hypoglycemia of infancy
OMIM 256450	-
Antigen peptide transporter 1 (APT1) (Peptide transporter TAP1) (Peptide transporter PSF1) (Peptide supply factor 1) (PSF-1) (Peptide transporter involved in antigen processing 1).	
OMIM 170260	TRANSPORTER, ATP-BINDING CASSETTE, MAJOR HISTOCOMPATIBILITY COMPLEX
Canalicular multispecific organic anion transporter 1 (Multidrug resistance-associated protein 2) (Canalicular multidrug resistance protein	
OMIM 601107	Dubin-Johnson syndrome
OMIM 237500	-
Spastin	
OMIM 182601	Spastic paraplegia-4
OMIM 604277	Spastic paraplegia-4
OMIM 182601	-
Cystic fibrosis transmembrane conductance regulator (CFTR) (cAMP- dependent chloride channel)	
OMIM 602421	Cystic fibrosis
OMIM 219700	Congenital bilateral absence of vas deferens
OMIM 277180	Sweat chloride elevation without CF ; {Pancreatitis, idiopathic} ; {Hypertrypsinemia, neonatal}
ATP-binding cassette, sub-family A, member 1 (ATP-binding cassette transporter 1) (ATP-binding cassette 1) (ABC-1) (Cholesterol efflux regulatory protein)	
OMIM 600046	Tangier disease
OMIM 205400	HDL deficiency, familial
OMIM 604091	-
Antigen peptide transporter 2 (APT2) (Peptide transporter TAP2) (Peptide transporter PSF2) (Peptide supply factor 2) (PSF-2) (Peptide transporter involved in antigen processing 2)	
OMIM 170261	Bare lymphocyte syndrome, type I, due to TAP2 deficiency

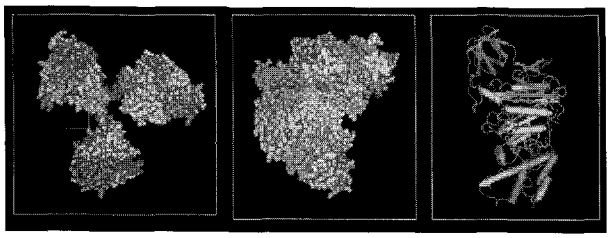

Fig. 3. Structures of 1BMF with highlighted [IPR003593] motif.

Conclusion

As we mentioned, the proteins in the world share motifs for evolutional or other biological benefits - they can save energy, time and resource for controlling and managing a variety of proteins. The hubs are structurally and functionally important in themselves and are also important in disease-related mutations. They are highly resistant to mutation to conserve their functions. But in the case of a rare mutation, on the position of hub can more easily cause fatal diseases. From an example of network, we knew that there could be several 'hub' motifs. The hub motifs act 'hubs' in the computer network system and are important in the aspect of structure and function of the proteins. As we have shown, the mutations (ex. missense mutation in OMIM) in these hub motifs can cause serious biological malfunction or disease to whole organism.

The major problem of this kind of network building is the redundancy control. The traditional association rule mining framework produces many redundant rules. So, we are trying to solve this problem and are planning to build non-redundant networks among motifs using that solution.

One of the other problems, there are reasonable threshold values. We used a minimum support 0.06% and a minimum confidence 50% threshold. However, various threshold values can be set according to the experimental purpose and many other forms of networks with different size and components.

Acknowledgments

This work was supported by Ministry of Science and Technology R&D Program. We would like to thank IBM Shard University Research (SUR) program, and CHUNG MoonSoul Center for Bio-Information and Bio-Electronics for providing computing facility used in this work.

References

Attwood, *et al.*, (2003). PRINTS and its automatic supplement, preprints. *Nucleic Acids Research* 31, 400-402.

Bateman, A. *et al.*, (2002). The Pfam Protein Families Database. *Nucleic Acids Research* 30, 276-280.

Deng, M. *et al.*, (2002). Inferring domain-domain interactions from protein-protein interactions. *Genomic research* 12, 1540-1548.

Flaquet, L. *et al.*, (2002). The PROSITE database, its status in 2002. *Nucleic Acid Research* 30, 235-238.

Han, J. and Kamber, M. (2001). Data Mining: Concepts and Techniques (Simon Fraser univ.) 225-228, 260-261.

Horng, J.-T. *et al.*, (2002). Study of Motif Correlation in Protein by Data Mining. *METMBS* 345-350.

Letunic, I. *et al.*, (2002). Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Research* 30, 242-244.

Mulder, N.J. *et al.*, (2003). The InterPro Database - 2003 brings increased coverage and new features. *Nucleic Acids Research* 31, 315-318.