MINING DOMAIN ASSOCIATION RULES FROM PROTEIN-PROTEIN INTERACTION DATA

Martin S.R. Paradesi[1,2], Liangjiang Wang[1], Susan J. Brown[1], William H. Hsu[2]

[1]Bioinformatics Center, Division of Biology, and [2]Department of Computing and Information Sciences, Kansas State University, Manhattan, KS 66506

ABSTRACT

Domains are conserved sequence regions in proteins. A protein often contains one to three domains, and the protein function may be inferred from the domain information. While many protein domains have been functionally characterized, the functions of some conserved sequence regions are still poorly understood. The objective of this work is to facilitate the functional annotation of domains using protein-protein interaction data. Our assumption is that if two or more domains co-occur in protein-protein interactions, these domains may be involved in the same biological process. In this work, we first investigate several association rule mining techniques for finding domain correlations. We then propose a new measure, *proximity*, for mining domain association rules. Finally, we discuss the potential uses of these rules for understanding domain functions.

INTRODUCTION

A protein domain is a functionally defined protein region. If a protein has multiple domains, then the combination of the domains determines the function of the protein. Protein-protein interactions refer to the association of proteins and the study of these associations. They play an important role because they determine which proteins take part in the same biological process. A large quantity of information about protein-protein interactions and protein domains has been accumulated over the years which is used in this study to find association rules among protein domains.

In this paper, we attempt to annotate unknown protein domain function by finding the association rules among protein domains by using the data from protein-protein interactions. The protein sequences are parsed for the functional domains using the HMMER tool. The *support-confidence-lift* structure of Association Rule Mining (ARM) along with a proposed measure *proximity* is applied to the protein-protein interaction data and the HMMER output information to find the correlation among functional domains of interacting proteins. *Lift* is used to measure how many times more often two protein domains occur together than expected if they were statistically independent, while *proximity* is used to find the probability of co-occurrence of two protein domains of an association rule in a single transaction (pair of interacting proteins).

Traditional ARM techniques are primarily used for finding association among products (or items) in business transactions as described by Agrawal *et al.* (1993). ARM techniques are used to discover the association between two sets of products such that the presence of some products in a particular

transaction implies that products from the other set are also present in the same transaction. Sarwar *et al.* (2000), formally define the fundamental concepts of ARM in the following manner. They denote a collection of m products $\{P_1,P_2,...,P_m\}$ by ρ. A transaction T is defined to be a subset of ρ consisting of products that are purchased together. An association rule between two sets of products X and Y, such that X,Y are subsets of ρ and $X \cap Y = \Phi$, states that the presence of products from the set Y are also present in T. Such an association rule is often denoted by $X \rightarrow Y$. The quality of association rules is commonly evaluated by looking at their *support* and *confidence*. The support s of a rule measures the occurrence frequency of the pattern in the rule, while the confidence c is the measure of the strength of implication. For a rule $X \rightarrow Y$, the support is measured by the fraction of transactions that contains both X and Y. Formally,

$$s = \frac{Number\ of\ transactions\ containing\ X \cup Y}{Number\ of\ transactions} \qquad (1)$$

In other words, support indicates that s% of transactions contain $X \cup Y$. For a rule , the confidence c states that c% of transactions that contain X also contain Y. Formally,

$$c = \frac{Number\ of\ transactions\ containing\ X \cup Y}{Number\ of\ transactions\ containing\ X} \qquad (2)$$

Thus, confidence is the conditional probability of seeing Y, given that we have seen X.

Brin *et al.* (1997), discovered that the confidence of two unrelated products might be sometimes quite high and eventually will generate a rule. This might occur while detecting the association rules between domains, because many domains are likely to occur with or without other items if only the support and confidence measures of an association rule are calculated. Hence, the *interest (lift)* is also calculated. Lift measures co-occurrence and not implication, in that it is completely symmetric. Formally,

$$lift = \frac{c(X \cup Y)}{s(Y)} \qquad (3)$$

PROPOSED MEASURE

The support, confidence and lift measures of ARM are inadequate to detect the association rule among protein domains because there might be few domains in a sequence and the number of protein-protein interactions (or transactions) is very large. Furthermore, in our case the association among domains is symmetric i.e., if there is a rule domain1 → domain2, then domain2 → domain1 will also hold true and have the same measures of the previous rule. Let us look at an example, by which the problem can be explained clearly. Let us assume that domainA and domainB co-occur in two interacting protein sequences. Let us also assume also that domainC and domainD co-occur in eight interacting protein sequences. The value of lift(domainA → domainB) in the first case is 0.5, while the value of lift(domainC → domainD) in the second case is 0.125. This is contradictory to the common intuition that the domains in the latter case

should be at least as highly correlated as the former case because they have more number of co-ocurring domains in a single sequence.

In order to overcome this shortcoming, we propose a new measure, *proximity,* which finds the correlation between two domains that occur in a single sequence. It is defined formally as,

$$\frac{(\textit{Number of transactions containing } X \cup Y)^2}{\textit{Number of transactions containing } X * \textit{Number of transactions containing } Y} \quad (4)$$

This measure gives a value between 0 and 1 and a higher value (closer to 1) indicates a higher correlation among the domains. This measure not only works well for the pair of domains which appear together in same interacting sequences, but also for the pair of domains which appear with other domains in same interacting sequences.

EXPERIMENT DESIGN

We retrieved the glocal alignment models of the Pfam Hidden Markov Model (HMM) library from the Pfam website and the protein-protein interaction data from the DIP database in FASTA format and XML format. The HMM file was converted to HMMER2 binary file using the *hmmconvert* tool in HMMER toolkit (Eddy, 2003). The HMMER2 binary file is given as an input to the *hmmcalibrate* tool that reads an HMM file, scores a large number of synthesized random sequences with it, fits an extreme value distribution (EVD) to the histogram of those scores, and updates the hmmfile to include the EVD parameters. The sequence file and the "calibrated" HMM file were given as input to the *hmmpfam* tool which looks for significantly similar sequence matches. The E-value cutoff for the per-sequence ranked hit list was set to 1.0.

The hits with E-value less than this cutoff are output into an output file in which there is a separate output report for each sequence in sequence file. This report consists of three sections: a ranked list of the best scoring HMMs, a list of the best scoring domains in order of their occurrence in the sequence, and alignments for all the best scoring domains. A sequence score may be higher than a domain score for the same sequence if there is more than one domain in the sequence; the sequence score takes into account all the domains. All sequences scoring above the -E and -T cutoffs are shown in the first list, then every domain found in this list is shown in the second list of domain hits.

The protein-protein interaction (PPI) data was downloaded from the DIP database. The data is stored in XML format and the proteins are represented as nodes and the interactions between proteins are represented as edges in the file. The file was parsed resulting in a list of 19,050 protein sequences and 54,511 protein-protein interactions.

Using the first section of the output file generated by HMMER along with the PPI data, we built an adjacency list containing the protein sequences along with the protein domains that were mapped to it and also found the number of unique domains that were present in the adjacency list. We calculated support, confidence and lift measures for ARM, assuming that an interacting sequence (i.e. a pair of sequences) is a transaction in the database and each domain is an itemset in the transaction. The support, confidence, lift and proximity were

calculated using the formulae as mentioned in the previous sections. The association rules consisting of the domains that occur in the same sequence were pruned because they do not provide valuable information about the correlation of the interacting proteins. The top-N association rules of protein domains with high proximity were collected for analysis. The experiment design is represented in Fig.1.

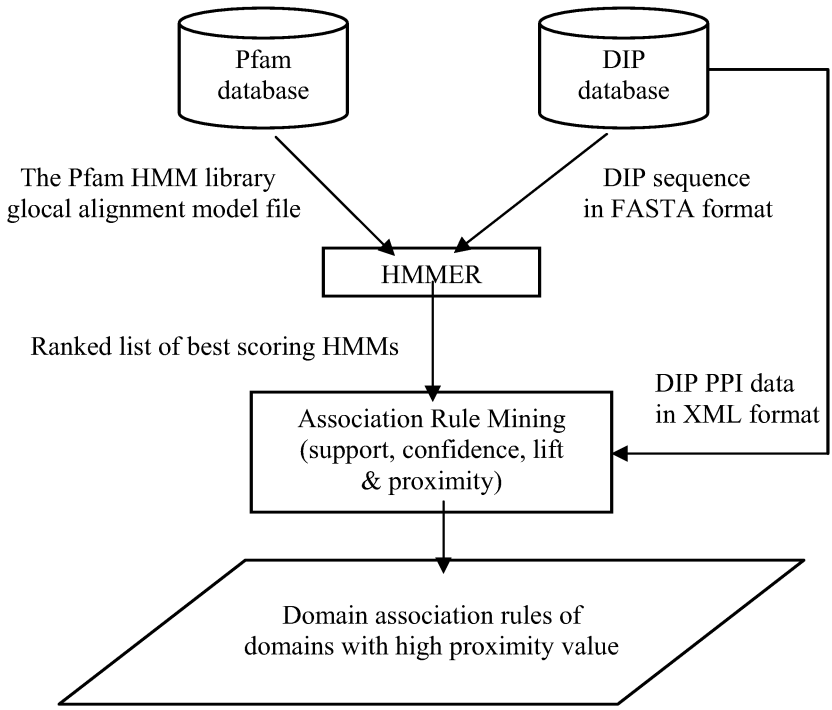

Figure 1. Framework for finding domain association rules using protein-protein interaction data.

RESULTS

Using the methodology as described in the experiment design section, the following results were observed. Applying *hmmpfam* on the DIP sequence file in FASTA format and the Pfam HMM library glocal alignment model file, 18,616 hits were detected which consist of 5,457 unique domains. On the application of ARM using minimal support = 0.000054, minimum confidence = 0.01, minimum lift = 2.0, and minimum proximity = 0.001, 320 rules were detected. These rules consist of highly correlated domains and few of them are as shown in Table 1.

Table 1. Comparison between *lift* and *proximity* values of the top eleven association rules

DomainA	DomainB	n(A)	n(B)	n(A UB)	*lift*	*prox-imity*
Phage_F	Phage_G	2	2	2	27255	1.0
Colicin_Pyocin	HNH	10	2	2	5451.1	0.2
Tissue_fac	Interferon	32	12	8	1135.6	0.166
Colicin_Pyocin	Cloacin	10	7	2	1557.4	0.057
CBF_beta	Runt	20	5	2	1090.2	0.04
IPK	FTCD_C	58	2	2	939.84	0.034
IPK	DUF1641	58	2	2	939.84	0.034
Mad3_BUB1_II	GbpC	13	9	2	931.81	0.034
Plug	TonB	23	23	4	412.18	0.03
MAT_Alpha1	STE	19	9	2	637.55	0.023
Exonuc_VII_S	DUF1291	50	4	2	545.11	0.02

Comparing the values of lift and proximity in Table 1, it is seen that the domains that are highly correlated have a higher proximity value, but do not follow any pattern in their lift values. It is observed that the association rule Phage_F → Phage_G has the highest proximity value because they occur together only in two interactions. According to the Pfam (Bateman *et al.* 2004) description, Phage_F is a family of proteins from single-stranded DNA bacteriophages and protein F is the major capsid component, sixty copies of which are found in the virion, while Phage_G is also a family of proteins from single-stranded DNA bacteriophages and the G protein is a major spike protein involved in attachment to the bacterial host cell and hence they are highly correlated protein domains. However, the remaining association rules do not have such high proximity values as they are less correlated when compared to the proximity between Phage_F and Phage_G protein domains. The eleventh row shows the association rule Exonuc_VII_S → DUF1291 and the proximity value of this rule is 0.02 which indicates that the domain DUF1291 may be involved in RNA degradation. It must also be noted that DUF1291 is not linked with Exonuc_VII_S in the same sequence. The sixth row and seventh row in the table above show the association rules between IPK → FTCD_C and IPK → DUF1641 with similar proximity values of 0.034. The knowledge gained using the two association rules is that DUF1641 may be involved in the Inositol polyphosphate kinase signaling pathway. Similarly, other domains with unknown functions can be looked up from the association rules that are generated by decreasing the proximity value, and then its correlated domains can be identified.

CONCLUSIONS & FUTURE WORK

In this paper, we have applied ARM to protein-protein interaction data to find association rules between domains. Because classical criteria for association rules - support, confidence and lift - do not sufficiently account for correlation, a

novel measure *proximity* is proposed which agrees well with the intution that protein domains that occur together in same sequence are highly correlated. The association rules among protein domains are detected and are sorted based on the value of their proximity values. One topic of future investigation is annotating more domains with unknown functions by considering the association rules with protein domains whose functions are known.

ACKNOWLEDGMENT

This work is supported by K-INBRE (NIH grant number P20 RR016475).

REFERENCES

Agrawal R., Imielinski T., Swami A., 1993, "Mining Association Rules between Sets of Items in Large Databases", ACM SIGMOD Record, 22(2):207—216

Bateman A., Coin L., Durbin R., Finn R.D., Hollich V., Griffiths-Jones S., Khanna A., Marshall M., Moxon S., Sonnhammer E.L.L., Studholme D.J., Yeats C., Eddy S.R., 2004, "The Pfam protein families database", Nucleic Acids Research, 32, D138--D141

Brin S., Motwani R., Ullman J.D., Tsur S., 1997, "Dynamic Itemset Counting and Implication Rules for Market Basket Data", Proceedings ACM SIGMOD International Conference on Management of Data, May 13-15, 1997, ACM Press (1997) 255—264

Eddy S.R., 1998, "Profile hidden Markov models", Bioinformatics, 14:755–763

Eddy S.R., 2003, "HMMER User's Guide, Biological sequence analysis using profile hidden Markov models", version 2.3.2

Oyama T., Kitano K., Satou K., Ito T., 2002, "Extraction of knowledge on protein-protein interaction by association rule discovery", Bioinformatics, 18(5):705—714

Salwinski L., Miller C.S., Smith A.J., Pettit F.K., Bowie J.U., Eisenberg D., 2004, "The Database of Interacting Proteins: 2004 update", Nucleic Acids Research, 32, D449--D451

Sarwar B.M., Karypis G., Konstan J.A., Riedl J., 2000, "Analysis of Recommendation Algorithms for E-Commerce", Proceedings of the 2nd ACM Conference on Electronic Commerce (EC-00), Minneapolis, MN, USA, pp. 158—167