Inferring Domain Combination Pattern and Its Biological Meaning via Association Rules

Suk-hoon Jung jsh@icu.ac.kr
Dong-soo Han dshan@icu.ac.kr
Sung-doke Lee sdlee@icu.ac.kr

School of Engineering, Information and Communications University, Munji-dong, Yusung, Deajeon 305-600, Korea

Keywords: domain combination, association rule, gene ontology

1 Introduction

All proteins consist of one or more domains with few exceptions. Domains are fundamental units of compact tree-dimensional structure, evolution, and hence the function. Researchers revealed that domains combined into proteins with limited repertoire [1]. Since proteins evolved through gene duplication, recombination, fusion and fission aiming toward specific functions, the fact that domain combination formation has limited rules is comprehensive. However, the biological meaning of domain combination has never sufficiently been researched except about pair-wise domain combination [1].

In this paper, we attempt to gain an overview of domain combination by studying domain combination patterns within proteins and analyzing them. We used modifications[3] of Association Rule[2] to find domain combination patterns whose member domains appear together frequently in the same protein. The data used for experiment are the 2586 proteins of *Saccharomyces cerevisiae* (baker's yeast) extracted from SWISS-PROT[4] which have domain information from Iterpro[5]. We also analyze functional annotation of proteins according to patterns obtained using Gene Ontology (GO)[6]. By this work, we verify that domain combination patterns, which might be sub parts of some proteins, are more functionally cohesive than the proteins what patterns belong to. It means that a domain combination pattern is assembled for specific functions and a protein might be several functional parts when having several disjoint domain combination patterns. These studies would be the sources of insights into domain combination and its biological meaning.

2 Method

2.1 Highly Affiliated Domain Combination Pattern

We used modifications[3] of Association Rules[2] to find domain patterns in proteins. Association Rules has widely been used in the field of data mining measuring the probability of appearance of items with prior condition of the other items in a set [2]. Since basic association rules are found with prior condition what is a part of items in a set, it should be modified to find domain combination pattern whose items are highly affiliated to each others. Therefore we use *h-confidence*[3], so we can capture the strength of domain combination association. Applying *h-confidence* and the concept of highly affiliated domain combination pattern, we obtained meaningful domain combinations whose members are highly associated with each others in proteins. From 2586 target data, we found 560 highly affiliated domain combination patterns with threshold = 0.5 and minimum support 0.0006 that means appearance more than twice. The patterns obtained cover 2258 proteins among 2586 ones

Definition 2.1 *The h-confidence of a pattern* $X = \{ d_1, d_2, \ldots. d_m\}$, *denoted as hconf(X), is a measure that reflects the overall affinity among domains within the pattern. This measure is defined as* $\min(conf(\{d_1\} \rightarrow \{d_2,\ldots d_m\}), conf(\{d_2, \rightarrow\{d_1,d_3\ldots, d_m\}\}, \ldots.., conf(\{d_m \rightarrow \{d_1,\ldots d_{m-1}\}))$, *where conf is the confidence of association rule.*

Definition 2.2 *A domain combination X is a highly affiliated domain combination pattern, when* $h\text{-}conf(X) \geq h_c$, *where* h_c *is a user-specified minimum threshold.*

2.2 Analyzing GO terms of Domains in Proteins

To research biological meaning of domain combination pattern, the analysis of GO terms[5] of domains would be a good approach. GO is the ontology for the feature of the gene products such as the protein and domain in three categories of 'cellular component', 'molecular function' and 'biological process'. We devised the *GO term Overlap Rate* (GOR). GOR_D is a representation of GO term Overlap Rate where D is set of domains and G_k is set of GO terms annotated to domain d_k which is an item of D. From the formula *(1)*, we measured the degree of cohesion of GO terms of domains in several protein groups according to the presence of patterns for three GO categories respectively.

$$GOR_D = \frac{|G_1 \cap G_2 \cap ... \cap G_k|}{|G_1 \cup G_2 \cup ... \cup G_k|} \qquad (1)$$

3 Result and Discussions

Figure 1 graphs the average GORs for several protein groups categorized according to the presence of highly affiliated domain combination patterns. The group *all proteins* are researched as a comparison group. For two GO term categories, molecular function and biological process, the average GORs of one-pattern-proteins are obviously higher than ones of two-pattern-proteins. Especially, the cohesions of biological function are strongly influenced by the number of patterns what proteins have, while the ones of molecular function are influenced by the presence of extra domains and the number of patterns. For the average GORs of cellular component, graph does not show major differences among the protein groups; it is comprehensible since GO terms for cellular component are decided by the physical locations of each protein.

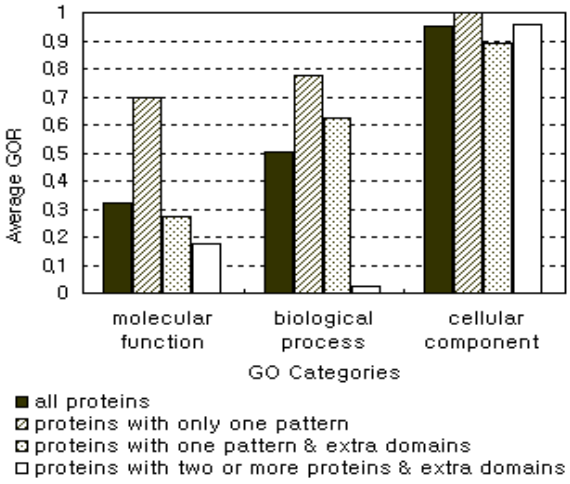

Figure 1: Average GORs of several protein Groups for three GO categories

Through this research, we found that domains tend to be combined into specific patterns whose elements are highly affiliated to each others. Also we verify that molecular function and biological process of GO term annotations are correlated with highly affiliated domain combination pattern. Therefore we proof that highly affiliated domain combination pattern found in proteins have biological meaning to be combined in the aspect of molecular function and biological process.

In the paper, we experiment only on baker's yeast data from SWISS-PROT, which could be thought not to be enough for conclusion. Furthermore, data set was trimmed against domain and GO term information, which causes diminishing the amount of data. Therefore applying the method to more huge data would produce concrete proof.

Acknowledgements

This work was supported by Korea Research Foundation Grant funded by the Korea Government (MOEHRD), Basic Research Promotion Fund (KRF-2005-050-D00013).

References

[1] Apic, G..& Gough, J. A., An insight into domain combinations, *Bioinformatics*, 17, S83-S89, 2001.
[2] Agrawal, R., T. Imielinski, and A. Swami, Mining association rules between sets of items in large databases, *Proc. Int. Conf. on Management of Data*, 207-216. 1993
[3] Hui Xiong, Pang-Ning Tan and Vipin Kumar, Mining Strong Affinity Association Patterns in Data Sets with Skewed Support Distribution, IEEE ICDM, 2003.
[4] http://us.expasy.org/sprot/
[5] http://www.ebi.ac.uk/interpro/
[6] http://www.geneontology.org/