Identification of Conserved Domain Combinations in S.cerevisiae Proteins

Suk Hoon Jung, Hee-Young Hur, Desok Kim and Dong-Soon Han
School of Engineering, Information and Communications University
P.O.Box 77, Yusong, Daejeon 305-600, Korea
Email: {jsh, hyerue, kimdesok, dshan}@icu.ac.kr

Abstract—In this paper, we propose a formulated method for the analysis of conserved domain combinations and report an overview of domain combinations by identifying domain patterns and analyzing their functional annotations. The proposed method measures co-occurrence frequency and mutual dependency of domains in a domain combination using association rules. The method is useful to estimate the meaningfulness of a given domain combination in terms of conservation. Using the method, we extracted domain patterns in S.cerevisiae proteins and investigated GO term annotations of the domains. According to the investigation, domains in S.cerevisiae proteins are turned out to form patterns in which the members of the patterns are highly affiliated to one another. Also, extracted patterns are revealed to have a tendency of being associated with molecular functions.

I. INTRODUCTION

All proteins consist of one or more domains with few exceptions. Domain is a conserved unit of compact three-dimensional structure and evolution [1], which carries specific function [2]. As conserved functional units, domains offer an abstract level at which the protein may be studied [3]. Therefore the detection of domains is one of the first steps toward assigning molecular functions to a protein. Consequently, domain-based protein function annotation resources are getting popular these days, and domains are frequently used as basic materials of computational prediction methods for protein function [4] [5] [6].

Although the domain works as a functional unit, the function of a domain should be considered in association with other neighbor domains in a protein due to polypeptide chain’s environmental sensitivity. Actually, neighbor domains of a domain are one of the most influential circumstances for the fragments’ folding and functions, and its experimental evidences were reported in several genes [7] [8] [9] [10] [11]. Domains in a protein might have relationship called domain interplay [11] or intra-domain communication [10]; domains help another domain in a protein or cooperate together for a function. Following this conception, several research groups started to use the notion of domain combination in computational methods [12] [13] [14], which take into account to the role of domain combination in the protein’s or individual domain’s function, and they obtained relatively good results. Therefore considering a protein just as a set of domains and investigating each domain independently might bring improper results in the study of protein functions.

When we consider that proteins have evolved toward specific target functions, domain may appear in association with other domains which have significant effects on the aimed functions. If so, those co-appearing domains in a protein, a domain combination, might be conserved throughout the evolution of the protein for the functional advantage. Several researches referred the conservation of domain combinations and revealed some evidences supporting the notion of domain combination [15] [16]. However, those researches do not explain the entire view of domain combination conservation phenomena in comprehensive manner because of the lack of systematical and biological analysis of domain combinations.

For the study of the conservation of domain combination, a method needs to be devised to measure the degree of conservations of each domain combination. Usually, protein sequence conservation has been evaluated by sequence alignment. Domain combination conservation should be evaluated in the same manner as well, yet, also, the relationships of domains in a combination should be considered.

In this paper, we fortify the notion of domain combination by identifying conserved domain combinations and analyzing biological meaning of the combinations. First, we develop a method for the analysis of conservation of domain combination. Association rule learning technique, which is widely accepted in data mining and treatment leaning, is used to identify conserved domain combinations. The method counts mutual dependencies between domains and co-occurrence frequencies in a combination. Unlike other previous researches, it systematically evaluates the significance of each domain combination irrespective of the number of members, versatility, and their continuity. The term of support denotes domain’s co-occurrence frequency, and all-confidence denotes the mutual dependency in a combination, respectively; support, and all-confidence are the terms used in association rules in general.

In this paper, a term, domain pattern, is coined to denote a part of a domain combination that is regarded as highly conserved domain combination. A domain pattern miner algorithm that identifies domain patterns from given protein is developed and domain patterns are extracted by the algorithm using predefined minimum support and all-confidence
threshold. For experiments, we used *Saccharomyces cerevisiae* (Baker’s yeast) proteins record from UniProt Knowledgebase [17] (release 9.3) and filter them using InterPro [18] domain information (release 13.1). Domain pattern miner has extracted 597 domain patterns from 340,459 possible domain combinations found from 3,457 Baker’s yeast proteins. The extracted patterns cover 90% of Baker’s yeast proteins.

To obtain biological meaning or members’ cooperation of conserved domain combinations, we have analyzed functional characteristics of domain patterns based on the information extracted from Gene Ontology (GO) terms annotated to each domain [19]. According to the study, domains in a domain pattern are revealed to have a tendency of having functional similarity that obeys our assumption about domain’s cooperation in conserved domain combination. Also, it is revealed that *all-confidence* have higher correlation with functional similarity than *support* does. Conventionally, *support* is mainly used for conserved domain combination identification. This indicates that the strategy of adopting *all-confidence* is promising to identify domain combinations contributing the same function. From the detail functional analysis, we drew the conclusion that conserved domain combination is correlative with molecular functions but not with biological process or cellular component. Therefore we can conclude that the domain pattern is the team for molecular functional collaboration and the collaboration is the reason why domain patterns have been assembled through the evolution.

II. RELATED WORK

A. Domains’ Interplay in a Combination

There are some studies on genes that report domain interplays or intra-domain communications within a protein. According to the studies, neighbor domains have high potential of indirectly influencing on working domain’s function irrespective of their inexplicit roles for target function [10] [11]. In some cases, domain is even revealed to directly influence neighbor domain’s function. In that case, the domain might explicitly take roles in enhancing, repressing or stabilizing the functions of neighbor domains [8]. Also domains might strengthen target function if they play the same functional role [7] [9].

As the study of domain combinations are focused on several remarkable genes, databases for those genes were prepared. KinG [14] and KinWeb [13] are the databases for Kinase adopted the notion of domain combination and its interplay. Kinase is a catalytic enzyme that transfers phosphate groups. The databases enable the retrieval of protein Kinase using domain combination. The reason of using domain combination is because, as they insist, the non-catalytic domains of protein Kinase are critical for the understanding of their biological roles. They reported that the association of the Kinase domain with other non-kinase domains within the same protein tightly regulates the activity of Kinase protein.

Those studies leads us to an idea that biological meaning of domain combination needs to be studied. Actually, PreSPI [12] was built up for the prediction of protein-protein interaction based on the notion of domain combination. PreSPI is the first trial in considering the possibility of the influence of adjacent domains on the domain-domain interaction. PreSPI achieved quite impressive prediction accuracy in the validation on Yeast proteins [12].

B. Domain Combination Conservation

Multi-domain proteins are assembled from existing combinations of domains with limited repertoire. For instance, about 90 percent of the roughly 600 enzymes in the small-molecule pathways of *Escherichia coli* are built from about 213 protein domain family [20].

The approach, which was analyzing multi-domain proteins in terms of pairs of domains adjacent to each other, has revealed several features of pair-wise domain combinations including that most domains families usually have a few major partners [15] [16]. Domain pairs or triplets that are conserved across domain architectures have been termed supra-domains by Vogel et al. [21]. They identified conserved domain combinations that are over-represented relative to occurrence of individual domains, and analyzed their function. The research is remarkable since they listed bunch of supra-domains even though the functional analysis was done by manually.

A graph approach was suggested to study the domain combinations. Ye and Godzik [22] developed a tool to visualize and compare domain co-occurrence graph. They showed that domain adjacency graph was highly clustered in some region while most parts of graph are sparsely connected. The observation indicates that domains have favorable partners to be combined in the same multi-domain protein. Another group showed similar results [23], but domain adjacency graphs contained direction information implying domain link direction on the sequence.

The researches clue us on domain combination phenomena in various way, co-occurrence and domain assembling versatility. However none of those elucidated the entire view of conserved domain combination caused from several limitation. They limited the number of domains in combinations, and domain assembling versatility or association was measured based on an interesting domain but not on all of members in a combination. Also, biological analysis was not sufficient due to the manual inspection.

III. CONSERVED DOMAIN COMBINATION IDENTIFICATION

A. Conserved Domain Combination Criteria

From proteins in an organism, we can extract huge amount of domain combination which may or may not have biological meaning. Domain combinations, appearing in several proteins within a genome, are likely to have evolved by gene duplication, so those are the result of evolutionary conservation of domains for some biological and functional advantage. Therefore, like the definition of conservation sequences, frequently found identical domain combination in an organism should be regarded as conserved and significant assembly.

In nature, several domains are abundant like as Kinase [20]. Proteins, what abundant domain belongs to, would generate
a frequent domain combination having abundant domain. Therefore this frequent combination, caused by member’s abundance, is not necessarily meaningful, and association analysis can overcome this problem.

Conserved domain combination may comprise domains that carry some biological meaning by members’ interplay, thus the association characteristics of domains should be understood. A domain has chemical and physical feature, so its roles in domain interplay are limited, and the domain would appear only in combinations in what it can take a role for target functions. Therefore the members of conserved combination should have dependencies on each other members. Also the dependency should be mutual since abundant domains compel dependency from minor domains. If every domain is mutually dependent on one another in the same combination, then we can say the combination is significant and conserved.

B. Domain Pattern Mining

Conserved domain combination is easily identified as domain pattern with data mining technique that fulfills two criteria, frequency and mutual dependency. We utilized two concepts in association rule running technique, support and all-confidence. Association rule running is widely used in the field of data mining and can represent significance and the strength of the itemset within the entire data. The notion support denotes the number of transactions that support an itemset [24]. Support for an itemset is defined as, in given transaction set, the fraction of transactions that contains all items of given itemset. In the context of this research, the item is domain, the itemset is given domain combination and the transaction is protein. As support counts the fraction of itemset, it is surely applicable to evaluate how frequently a domain combination occurs or domains occur together. Support corresponds to statistical significance, so motivation for support constraint comes from the fact that we are interested only in frequent appearing domain combination above predefined minimum support. If the support of a domain combination is not large enough, the combination is not thought to be conserved and not worth consideration.

Definition 1: Support of $dc$, a domain combination, is

$$\text{support}(dc) = \frac{|\{p|p \in P \land dc \subset p\}|}{|P|}$$

where $p$ is a protein in proteome $P$

Confidence measures the strength of association within itemset [24]. In the context of proteins and domains, an association rule is of the form $X \Rightarrow Y$, which means the presence of domain set $X$ implies the presence of domain set $Y$ in the same protein. The confidence of the association rule $X \Rightarrow Y$ is written as $\text{conf}(X \Rightarrow Y)$ as defined by Definition 2 or simply calculated by equation 1.

Definition 2: Confidence of $X \Rightarrow Y$ is

$$\text{conf}(X \Rightarrow Y) = \frac{|\{p|p \in P \land X \cup Y \subset p\}|}{|\{p|p \in P \land X \subset p\}|} = \frac{\text{support}(X \cup Y)}{\text{support}(X)} \quad (1)$$

All-confidence is a measure of the interestingness of an association, whose result value can be regarded as a degree of mutual dependency within an itemset [25]. All-confidence value is the minimum of the confidence values of all rules that can be produced from target itemset. Also, with the predefined minimum threshold, an association is deemed interesting if it has an All-confidence greater than the threshold. This indicates that there is a dependency among all of the items in the association. For example, if the All-confidence value is one, the any subset of given itemset would imply the remaining items with a confidence of 100 percent. Certainly, in that case, there is a high degree of mutual dependency among the items in given itemset.

Since basic confidence is measured with prior antecedent condition and item orientation, it could not applicable for domain combination when we are interested in combination but not in certain domain in a combination. In contrast to the confidence, all-confidence is useful measure of mutual dependency within a combination regardless of orientation of domains. Therefore it can surely be applied to measuring the strength of the domain combination.

Definition 3: The all-confidence of a domain combination, $dc$, is

$$all - \text{conf}(dc) = \frac{|\{p|p \in P \land dc \subset p\}|}{\text{MAX}\{|i|\forall l(l \in \text{PowerSet}(dc) \land l \neq \phi \land l \neq dc \land i = |\{p|p \in P \land l \subset p\}|\}}$$

Domains in a conserved domain combination should be associated with and dependent on one another, also they should be appeared together frequently. Therefore domain pattern mining with predefined minimum support and all-confidence threshold must be promising for conserved domain combination identification.

Even though a domain combination has values that exceed predefined constraints of support and all-confidence, it could be useless as a domain pattern. Some domain combination has superset with the same support, and that means the subset occurs only when the superset does. In that case, subset is meaningless, or we can not measure the meaning of subset with given protein data. Therefore, those domain combinations should be trimmed before the analysis, so we defined maximal property for the trimming and used to define domain patterns.

Definition 4: A domain combination $X$ has maximal property if no superset of this combination has the same or greater support.

Definition 5: A domain combination $X$ is a domain pattern if it has maximal property, $\text{supp}(X) > s_c$ and all-$\text{conf}(X) > a_c$ where $s_c$ and $a_c$ are predefined thresholds.

According to predefined minimum support and all-confidence threshold, various domain patterns sets could be defined with different conservation degree. Therefore minimum support and all-confidence threshold should be defined after examining the characteristic of target organism or protein set.
Briefly, the process of searching domain combinations can be viewed as the generation of a level-wise pattern tree. We adopted apriori algorithm [26] to domain pattern miner for generation of domain combinations; generated combinations are candidates for domain patterns and will be pruned based on maximal property, minimum support and all-confidence threshold.

C. Functional Analysis of Domain Combination

We introduce our strategy to measure functional similarity among given domain combination exploiting GO term information annotated to each member domain. In this research, we use a term, Inner Functional Similarity (IFS), denoting whether members of a domain combination are devoting the similar function or not.

GO is the ontology for the feature of the gene products. GO is a set of structured vocabularies organized in a rooted directed acyclic graph (DAG), describing attributes of proteins, domains or RNA in three categories of cellular component, biological process and molecular function. Each of GO categories should be analyzed respectively as they have different biological meaning.

The functional similarity of two GO terms must be considered with hierarchical manner as GO is ontology. Therefore, we adopted FuSSiMeG function [27] to investigate similarity of two GO terms. FuSSiMeG, which is a tool computing the semantic similarity between two GO terms, exploits Jiang and Conrath’s semantic similarity measure that provides the best result overall [27]. This semantic similarity measure is a hybrid approach; it combines information content and conceptual distance with parameters that control the degree of each factor’s contribution.

Since FuSSiMeG generates similarity value for two GO terms, it should be extended into IFS for observing functional relationship of multiple GO terms annotated to each domain in a combination which might have more than two domains; extension is shown in Equation 2

\[
IFS(G) = \frac{\text{Sum}(\{s|\forall g_i, g_j (g_i \in G \land g_j \in G \land i < j \land s = \text{FuSSiMeG}(g_i, g_j)\}})\right)\right)}{\left(\frac{|G| \cdot (|G| - 1)}{2}\right)}
\]

(2)

Let G is a GO term set whose members are annotated to each domain in a combination. First, all possible GO term pairs from are generated from G, and similarity values for each pair are reserved by FuSSiMeG. Then, we sum up all reserved similarity values and divide the sum by the number of pairs. In a nutshell, IFS is the average of FuSSiMeG values of possible GO term pairs of a domain combination.

IV. RESULTS AND DISCUSSION

In this section, we illustrate domain pattern mining and functional analysis procedures using S.cerevisiae proteome, Baker’s yeast proteins. First of all, all domain combinations in Baker’s yeast proteome are evaluated by proposed method. Then we investigate relations between measured values and functional features of each domain combination. Finally, we approve parts of whole domain combinations as domain patterns, and then report there functional features.

A. Domain Pattern Candidates

Eukaryote is well known to have more complex protein structure than prokaryote [28]. Among eukaryotes, Baker’s yeast is one of the well studied species. Therefore Baker’s yeast proteome would be adequate data source to reveal domain combination phenomena and to examine domain pattern approach. We used 7,449 Baker’s yeast proteins recorded in UniProt Knowledgebase [17] (release 9.3), and it was filtered using InterPro [18] (release 13.1) domain information. Information of used data is shown in Table I.

First, we generated 340,459 domain combinations that were found at least once from in Baker’s yeast proteins. Pattern candidates were also generated applying maximal property, and it was revealed that only 1,758 maximal combinations are remained from 340,459 domain combinations. That dramatic decrease of domain combination is the evidence that multi-domain proteins were assembled from existing combinations of domains with limited repertoire.

B. IFS Distribution

Using generated maximal domain combination, we investigated functional similarity tendency of domain members in a combination against support and all-confidence values. We applied functional similarity measure IFS, and IFS were performed for three GO term categories respectively. Since GO term information is insufficient, only parts of domain combinations are measurable (Table II).

Figure 1 plots the distribution of IFS against all-confidence for molecular function category. The distribution is regressing to IFS 100 as all-confidence is getting greater, but dots tend to be distributed over low IFS area at low all-confidence. The graph shows that IFS has upward tendency vividly as all-confidence is getting close to one. If we compare that with the Figure 2, it becomes obvious that IFS for molecular function of domain combination is related to all-confidence rather than to support; the regression line shows that IFS and support are fairly correlated, but distribution is not obvious.

In another view point, we can see that the distribution of Figure 1 is not uniformed, so we need to observe keenly. Dots
on Figure 1 could categorized into two parts for convenience sake, ones at IFS 100 and ones that do not hit at IFS 100. Dots at IFS 100 spread over all confidence although they are biased in favor of high confidence. Dots not at IFS 100 seem to spread low IFS. We infer from the observation that the distribution might reflect several types of domain-interplay. If domains in a combination prepare physical and geological backups that implicitly enhance or repress working domain’s function [10] [11], the domain combination does not need to have the same or similar functional annotations. Contrarily, if they cooperate explicitly for target functions [7] [8] [9], members in a combination should have the same function. This hypothesis is probable, but more researches are necessary for the proof.

Figure 3 shows the distribution of IFS for biological process GO category against all-confidence, and Figure 4 shows the one against support. Both graphs show that most of IFSs are ranked 100 while a few dots are plotted at low IFS.

The relations in graphs are not obvious, so necessity of Pearson’s correlation test comes up with that reason. Table II contains the numerical correlation values for each cases calculated using Pearson’s correlation test. The correlation coefficient between IFS of molecular function and all-confidence is 0.464, and the one with support is 0.205. Therefore we can say that IFS of molecular function is more related with all-confidence than support. As significant values p for correlation coefficients of molecular function are 0, which are smaller than significant level 0.01, the correlation coefficients calculated are statistically acceptable.

For IFS of biological process, the correlation coefficient with all-confidence is 0.268, and the one with support is 0.129. However they could not be comparable since correlation coefficient with support are not accepted because of the significant value p which is grater than significant level 0.01. Pearson’s correlation test for cellular component is also not statistically acceptable as significant values p are grater than
significant level 0.01.

As, conventionally, frequently appeared homology sequence has been regarded having significant function, domain patterns with high support and all-confidence are also expected to have functional significance. Moreover, functional tendency is more related to all-confidence than support, which implies that proposed approach adopting all-confidence is more promising to identify conserved domain combination.

### C. Domain Pattern in Yeast Protein

The next step of domain pattern inference should be determining minimum support $s_c$ and all-confidence threshold $a_c$. Those constraints let us specify conservation degree of domain patterns. $s_c$ and $a_c$ could be determined arbitrary, but it is recommended to choose them after observation of data. To obtain biologically meaningful and enough domain pattern, we choose 0.3 for all-confidence threshold $a_c$ and 0.00057 for minimum support $s_c$ regarding to distributions. With specified $s_c$ and $a_c$, domain pattern miner generated 597 domain patterns from 1,758 domain pattern candidates in Baker’s yeast proteins. Those obtained patterns cover 3,088 proteins among 3,457 given proteins.

Domain pattern candidates, which are domain combinations having maximal property, were categorized two groups, Pattern and None pattern. Those groups should have biological differences if our domain pattern inferring method worked well. We analyzed IFS for each domain pattern in Pattern group and each of None group. The numbers of measurable targets are shown in Table III.

The result is shown in a box plot, Figure 7. For molecular function, domain pattern candidates seem to be categorized well. The median of Pattern group is 100 while the one of None group is around 50 where fourth quartile of Pattern is ranked on. However, IFS values of group None are not regular. It might be caused by inadequate predefined thresholds or lacks of Yeast protein data.

For biological process, the medians of Pattern and None groups are not surely distinguished; only third quartile of Pattern group exceeds the one of none group while first, second and fourth quartiles are the same.

For cellular component, the combinations in Pattern group are functionally the same within domain members. The median, maximum and even minimum values are graded at IFS 100. None group also have high IFS values; almost all of IFS are ranked at 100 except few extremal values. Those facts were already forecasted in Figure 5 and Figure 6 as domain combination had high IFS values even at low degree of support and all-confidence.

We reserved the results of T-Tests of Pattern and None groups in each aspect of three GO term categories for neutralization of the differences between them. Since only significant value of molecular function is smaller than 0.01, the differences between Pattern and None for molecular function are statically proved. Therefore, domain patterns obtained, which is categorized in Pattern, would surely be more molecular functionally similar within members than domain combination that were not recognized as domain patterns. The power of domain pattern approach seems to work mainly on GO term category molecular function. From those results, we can infer that conserved domain combination takes roles of small functions like molecular function rather than cellular component.

### V. CONCLUSION

In this research, we developed a formulated method for identifying conserved domain combination using support and all-confidence. Proposed method enables us to explain the domain combination conservation quantitatively, so domain combinations can be listed or sorted according to their values.
Using the method, we studied domain combinations by measuring conservation degree and analyzing functional characteristics. We obtained 597 conserved domain combinations, defined as domain patterns, whose members frequently appear together and mutually dependent on one another. The experiments applying IFS (Inner Functional Similarity) measurement showed that domain pattern has correlation with molecular function of GO term category. This provides us some clues in explaining the reason why conserved domain combinations were assembled through evolution. That is, domains form a team in some case for constructing specific molecular functions. The results support that proposed method is better than conventional methods in identifying conserved domain combinations in which domains devote the same function. This is because the method adopted mutual dependency of domains within combination in measuring conservation degree.

Consequently, when looking at molecular function of proteins and using proteins as the functional objects, investigation of conserved domain combination deserves to be considered rather than examining single domain separately. Besides well filtered domain patterns can provide clues in various biological findings such as functional prediction or domain interplay discovery.

In the future, we are planning to make other findings for domain interplay or to predict domain’s functional collaboration in a protein by concentrating more effort on functional analysis of domain combination. As reported in previous researches on domain interplay, there are diverse forms of collaborations among domains to build a specific function. Thus, we need to build more sophisticated models to explain functional meanings of domain combinations.

REFERENCES


