Conserved Domain Combination Identification in Human Proteins

Suk Hoon Jung, Desok Kim, Dong-Soo Han *
School of Engineering, Information and Communications University
,Yusong, Daejeon, Korea
{jsh, kimdesok, dshan}@icu.ac.kr

Abstract

In this paper, we propose a method for the analysis of conservation of domain combinations in proteins. Using the method, we extract conserved domain combinations in Homo sapience proteome and examine their GO term annotations in order to understand the co-evolution of domains in a proteome. Unlike conventional methods, which use co-occurrence frequency for evolutionary analysis of domains, the proposed method measures mutual dependency of domains in proteins as well. According to the results, domains in Homo sapience proteome turn out to form patterns whose members are highly affiliated to one another. Besides, GO term analysis shows that extracted patterns have a tendency of being associated with molecular functions, and molecular functions are more related with mutual dependency than co-occurrence frequency of domains. Those results indicate that the proposed method adopting mutual dependency outperforms conventional methods in terms of finding domain combinations conserved through evolution for molecular functional collaboration.

1 Introduction

Domain is a conserved unit of compact three-dimensional structure and evolution [13], which carries specific function [14]. As conserved functional units, domains offer an abstract level at which the protein may be studied [10], so the detection of domains is one of the first steps toward assigning functions to a protein. Therefore, domain-based protein function annotation resources are getting popular these days [6], and accumulated domain resources are widely utilized in laboratory and computational methods for protein function discovery. [11].

Although a domain has its own function, the function of a domain should be considered in association with neighbor domains due to environmental sensitivity of polypeptide chain. Actually, neighbor domains are one of the most influential circumstances for the domain folding and function, and its experimental evidences were reported in several genes. Domains in a protein might have relationship called *domain interplay* [17] or *intra-domain communication* [3]. A domain might indirectly influences on neighbor domains in a protein [3] [17], explicitly takes a role in enhancing, repressing or stabilizing the functions of neighbors [4], or strengthen target function with neighbors playing the same functional role [9]. With this conception, several research groups started to use the notion of domain combination in protein function prediction methods considering neighbor effect [12] [15], so that they obtained relatively good results. Therefore considering a protein just as a set of domains and studying each domain independently might bring improper results in the study of protein function.

When we consider that proteins have evolved toward specific functions, a 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, must be conserved through protein evolution for functional interplay. Several researches clue us on conserved domain combination phenomena in various way, co-occurrence and domain assembling versatility, and give some evidences supporting the notion of conserved domain combination [2] [19]. However none of those elucidates the phenomena of conserved domain combination, as they mainly explain only the master and servant relationship between two domains . They limit the number of domains in combinations, and domain assembling versatility or association is measured based on an interesting domain but not on all of members in a combination. Besides, biological analysis is not sufficient due to the manual inspection.

For the study of the conservation of domain combination, a systematic method needs to be devised to measure the conservation degree of each domain combination. Usually, protein sequence conservation has been evaluated by sequence alignment. Domain combination conservation should be evaluated in the similar manner as well, yet also the rela-

*Corresponding author: School of Engineering, Information and Communications University P.O.Box 77, Yusong, Daejeon 305-600, Korea (telephone: +82-042-866-6290; e-mail: dshan@icu.ac.kr)

978-0-7695-3096-3/08 $25.00 © 2008 IEEE
DOI 10.1109/WAINA.2008.228

tionships 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 functional relations among domains in a combination. First, we develop a quantitative method for the analysis of conservation of domain combination as applying two concepts of association rule learning technique that is widely used in data mining and treatment leaning. The method measures mutual dependency between domains and co-occurrence frequency in a proteome. Unlike other previous researches, it systematically evaluates the significance of each domain combination in terms of conservation, 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. The term of domain pattern is coined to denote a domain combination that is regarded to be highly conserved through evolution. A domain pattern miner algorithm extracts domain patterns using predefined *minimum support* and *all-confidence* threshold.

From the experiments, we used record from UniProt Knowledgebase [7] (release 9.3) and refined them against InterPro [6] domain information (release 13.1). *Domain pattern miner* has extracted 681 domain patterns from 39,563 *Homo sapience (human)* proteins. Four fifth of human proteins have one or more domain patterns.

In order to elucidate biological roles or members' cooperations of conserved domain combinations, we have analyzed domain patterns based on Gene Ontology (GO) terms annotated to each domain [5]. The GO term analysis reveals that domains in a pattern have a tendency of having functional similarity, which obeys our assumption about domains' cooperation in conserved domain combination. Also, it is revealed that *all-confidence* have higher correlation with functional similarity than *support* does; *support* is conventionally used for conserved domain combination identification. These facts indicate that the strategy adopting mutual dependency outperforms methods using only co-occurrence frequency in identifying domain combinations contributing to the same function. Besides, The comparison among the results of analysis on three GO term categories shows that the domain patterns are correlative with molecular function but not with biological process or cellular component. Therefore, we conclude that the conserved domain combination is the team for molecular functional collaboration and the collaboration is one of the reasons why the combination have been assembled through the evolution.

2 Materials and Methods

2.1 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 [18]. 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.

2.2 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 [1]. *Support* for an itemset is defined as, in given transaction set, the fraction of transactions that contains all items of given itemset. In the case 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*

$$supp(dc) = \frac{|\{p|p \in P \wedge dc \subset p\}|}{|P|}$$

where p is a protein in proteome P

Confidence measures the strength of association within itemset [1]. In the context of proteins and domains, an association rule is of the form *X* ⇒ *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* ⇒ *Y* is written as *conf(X* ⇒ *Y)* as defined by Definition 2.

Definition 2 *Confidence of X* ⇒ *Y is*

$$conf(X \Rightarrow Y) = \frac{|\{p|p \in P \wedge X \cup Y \subset p\}|}{|\{p|p \in P \wedge X \subset p\}|}$$

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 [16]. *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 threshold. This indicates that there is a dependency among all of the items in the association.

Since basic *confidence* is measured with prior antecedent condition and item orientation, it could not applicable for domain combination when we are not interested in certain domain in a combination. In contrast to the *confidence*, *all-confidence* is useful measure of mutual dependency within an itemset regardless of orientation of items. 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*

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

Domains in a conserved domain combination should be associated with and dependent on one another, also they should be co-occurred frequently. Therefore domain pattern mining with predefined *minimum support* and *all-confidence thresholds* 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 has no meaning, or we can not measure the meaning of subset with given protein data. Therefore, those domain combinations should be trimmed before observing *support* and *all-confidence*. These characteristic is defined as *maximal property* and used to define domain pattern.

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, and supp(X) > s_c all-conf(X) > a_c where s_c and a_c are predefined minimum thresholds.*

According to predefined *minimum support* and all-confidence thresholds, of course, various domain pattern sets could be defined. Therefore *minimum support* and *all-confidence threshold* should be defined after examining the characteristic of target organism or protein set.

2.3 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, Internal Function 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 [8] 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 [8].

Since FuSSiMeG generates similarity value for two GO terms, it should be extended into *IFS* for observing functional relationship of multiple GO terms annotated domain

Table 1. Experimental and Generated Data

Data	Number
protein	70490
domain	2334
protein with domain information	39563
multi-domain protein	24607
domain combination found in proteins	4984605
maximal domain combination	3593

combination which might have more than two domains; extension is shown in Equation 1. In a nutshell, *IFS* is the average of FuSSiMeg values of possible GO term pairs of a domain combination.

$$IFS(G) = \frac{Sum(\{s|\forall g_i g_j(g_i \in G \wedge g_j \in G \wedge i < j \wedge s = FuSSiMeG(g_i, g_j))\})}{\left(\frac{|G| \cdot (|G| - 1)}{2}\right)} \quad (1)$$

3 Results

In this section, we illustrate domain pattern mining and functional analysis procedures using *Homo sapience* proteome, human proteins. First of all, all domain combinations in human proteome were evaluated by proposed method. Then we investigated relations between measured values and functional features of each domain combination. Finally, we approved parts of domain combination as domain pattern, then analyze its functional features.

3.1 Domain Pattern Candidates

We used 70,490 human proteins recorded in UniProt Knowledgebase [7](release 9.3), and it was filtered against InterPro [6](release 13.1) domain information. Information of used data is shown in Table 1.

First, we generated 4,984,605 domain combinations that is found from at least one protein in human proteins. Pattern candidates were also generated applying *maximal property*, and it was revealed that only 3,593 maximal combinations are remained from 4,984,605 domain combination. That dramatic decrease of domain combination is an evidence that multi-domain proteins were assembled from existing combinations of domains with limited repertoire. Using 3,593 maximal combinations, we can form any human proteins, so the maximal domain combination is another type of the unit of the protein.

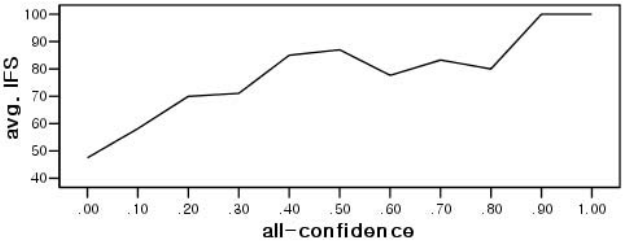

Figure 1. Average IFS for Molecular Function against all-confidence

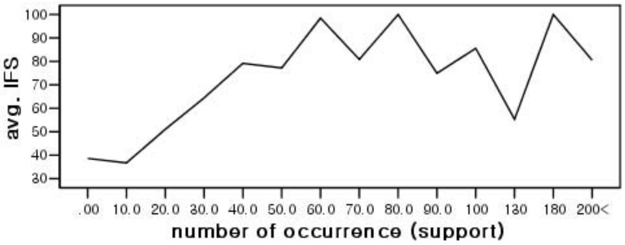

Figure 2. Average IFS for Molecular Function against support

3.2 IFS Tendency

Using generated *maximal domain combination*, we investigated functional similarity tendency of domain members in a combination according to *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 2).

Table 2 contains the numerical correlation values for each cases calculated using Pearson's correlation test. The correlation coefficient between molecular function and *all-confidence* is 0.376, and the one with *support* is 0.250. 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.

However the correlations for biological process and cellular component could not be comparable since correlation coefficient with *support* are not statistically accepted because of the significant value *p* which is grater than significant level 0.01.

Figure 1 plots the relationship between all-confidence and average *IFS* in the aspect of molecular function category. As the line go upward, the graph shows that average *IFS* has upward tendency vividly as all-confidence get-

Table 2. Pearson's correlation test result

IFS	All-confidence		Support		# data
	r	*p*	*r*	*p*	used
IFS_{mol}	0.376	0.000	0.250	0.000	334
IFS_{bio}	0.227	0.000	0.177	0.007	230
IFS_{cell}	0.250	0.0836	0.209	0.149	49

r: correlation coefficient
p: significant value
mol: molecular function go term category
bio: biological process go term category
cel: cellular component go term category

ting close to one. If we compare that with the Figure 2, it becomes obvious that average *IFS* of domain pattern candidates for molecular function is related to all-confidence rather than to support. The line is fluctuated in high support region of the graph Figure 2 which plot the relationship between support and average *IFS*. When the number of the occurrence is relatively small, *IFS* line seems to be increased against support. However, upward tendency of *IFS* is disappeared and seems not to be stable where the number of the occurrence is more than a hundred.

As, conventionally, conserved 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.

3.3 Domain Pattern in Human 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 0.2 for *all-confidence threshold* a_c and 0.0003 for *minimum support* s_c after observing data distributions. With specified s_c and a_c, *domain pattern miner* generated 681 domain patterns from 3,593 maximal combinations in Human proteins. Those obtained patterns cover 32172 proteins among 39563 given proteins.

Now domain pattern candidates, which are domain combinations having maximal property, were categorized two groups, *Pattern* and *None* groups. 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. Due to lack of GO term information, We approve combination to measure *IFS* if more than four fifth of domains have GO terms.

Table 3. T-test result

		#Case	Avg.	SD	*t*	*p*
molecular function	*pattern*	82	82.0	29.4	6.0	0.000
	none	252	51.8	35.0		
biological process	*pattern*	39	74.7	33.6	3.6	0.000
	none	191	68.6	35		
cellular component	*pattern*	14	100	0	0.9	0.319
	none	35	92.9	33.2		

Avg.: Average of functional similarity.
SD: Standard deviation
t: *t*-Value
p: significant value

The number of measurable targets are shown in Table 3.

The result is shown as a box plot in Figure 3. 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 40. However, *IFS* values of group *None* are not regular. It might be caused by predefined thresholds or lacks of human protein data.

For biological process, the *IFS* of *Pattern* and *None* groups are not surely distinguished. the of *Pattern* is 100, but third and fourth quartiles of *Pattern* are below the median of *None*. For cellular component, *Pattern* and *None* group rarely have differences almost all of *IFS* are ranked at 100 except few extremal values.

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.

4 Conclusion

In this research, we developed a systematic method for identifying conserved domain combination in human proteins 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 degrees and analyzing functional relation among domains in a combination. We obtained 681 conserved domain combinations, defined as domain pat-

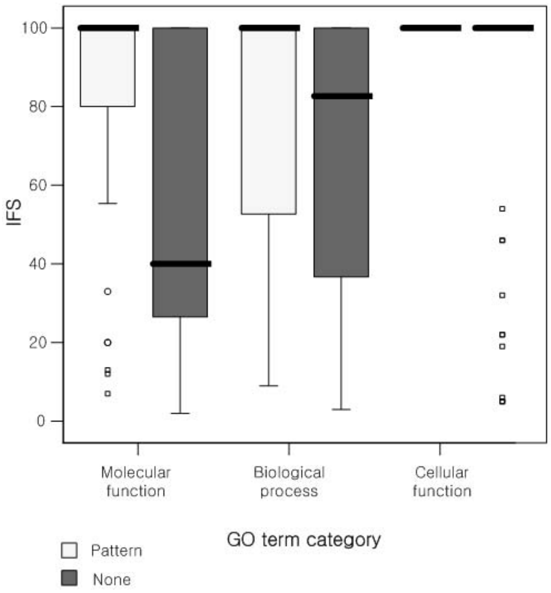

Figure 3. IFS Distribution for *Pattern* and *None* groups

terns, whose members frequently appear together and are mutually dependent on one another. The analysis applying *IFS* (Inner Functional Similarity) measurement shows that domain patterns have correlation with molecular function of GO term category. The results explain one of the reasons why conserved domain combinations were assembled through protein evolution; that is, domains form a team for constructing specific molecular function when the function require collaboration among domains. Also, the results support that proposed method exceeds conventional methods in identifying conserved domain combinations in which members contribute to constructing target function. This is because the method employ mutual dependency of domains within combination in measuring conservation degree.

Consequently, when looking at molecular function of proteins, investigation of 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.

5 Acknowledgement

This research was financially supported by the Ministry of Commerce, Industry and Energy (MOCIE) and Korea Industrial Technology Foundation (KOTEF) through the Human Recourse Training Project For Reginal Innovation

References

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. *Proceedings of the ACM SIGMOD International Conference*, (Management of Data):207–216, 1993.

[2] G. Apic, J. Gough, and S. Teichmann. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J. Mol. Biol.*, 310:311–325, 2001.

[3] J. Brodie and I. J. McEwan. ntra-domain communication between the n-terminal and dna-binding domains of the androgen receptor: modulation of androgen response element dna binding. *Journal of Molecular Endocrinology*, 34:603–615, 2005.

[4] C. C. T. Chen and A. Shyu. Interplay of two functionally and structurally distinct domains of the c-fos au-rich element specifies its mrna-destabilizing function. *Mol. Cell Biol.*, 14(1):416–426, 1994.

[5] T. G. O. Consortium. Gene ontology: tool for the unification of biolog. *Nature Genet.*, 25:25–29, 2000.

[6] U. Consortium1. The universal protein resource (uniprot). *Nucleic Acids Res.*, 35(Database issue):D224–8, Jan 2007.

[7] U. Consortium1. The universal protein resource (uniprot). *Nucleic Acids Res.*, 35(Database issue):D193–7, Jan 2007.

[8] F. Couto, M. Silva, and P. Coutinho. mplementation of a functional semantic similarity measure between gene-products. *Department of Informatics*, pages 3–29, 2003.

[9] P. Devarajan, D. A. Scaramuzzino, and J. S. Morrow. Ankyrin binds to two distinct cytoplasmic domains of na,k-atpase alpha subunit. volume 91, pages 2965–2969, 1994.

[10] M. B. A. et al. Cdd: a curated entrez database of conserved domain alignments. *Nucleic Acids Res.*, 31:383–387, 2003.

[11] R. A. et al. The interpro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, 29:37–40, 2001.

[12] D.-S. Han, H.-S. Kim, W.-H. Jang, S.-D. Lee, and J.-K. Suh. Prespi: a domain combination based prediction system for protein-protein interaction. *Nucleic Acids Res.*, 32:6312–6320, 2004.

[13] T. Hubbard, A. Murzin, S. Brenner, and C. Chothia. a structural classification of proteins database. *Nucleic Acids Res.*, 25:236–239, 1997.

[14] M. Hurles. Gene duplication: the genomic trade in spare parts. *PloS. Biol.*, 2:E206, 2004.

[15] A. Krupa, K. Abhinandan, and N. S. and. A database of protein kinases in genomes. *Nucleic Acids Res.*, 32(Database issue):D153–D155, 2004.

[16] E. R. Omiecinski. Alternative interest measures for mining associations in databases. volume 15, pages 57–69, 2003.

[17] N. B. E. Ronne and K. Dano. Domain interplay in the urokinase receptor. *J. Biol. Chem.*, 217(37):22885–22894, 1996.

[18] S. Teichmann, S. Rison, J. Thornton, M. Riley, J. Gough, and C. Chothia. Small-molecule metabolism: an enzyme mosaic. *Biotechnol*, 19:482–486, 2001.

[19] C. Vogel, C. Berzuini, M. Bashton, J. Gough, and S. Teichmann. Supra-domains: evolutionary units larger than single protein domains. *J. Mol. Biol.*, 336(3):809–823, 2004.