Enrichment of Protein Utilization in Protein-Protein Interaction Prediction by Adjusting e-value of InterProScan and Using Gene Ontology

Woo-Hyuk Jang[1] **Hee-Young Hur**[1] **Dong-Soo Han**[1]

torajim@icu.ac.kr hyerue@icu.ac.kr dshan@icu.ac.kr

[1] School of Engineering, Information and Communications University, 119, Munjiro, Yuseong-gu, Daejeon, 305-714, Korea

Abstract

Among the computational prediction methods, domain based protein interaction prediction approaches are getting popular. However, since conventional domain based approaches use only domain to identify proteins, two different proteins cannot be distinguished if they share same domains. Moreover, proteins which have no domain information cannot be used in the domain approaches, so not all interaction pairs are available. In this paper, two different ideas are proposed; first one is an increasing domains by adjusting e-value of InterProScan which extracts domains from a raw sequence, and the other is a utilizing gene ontology (GO) which describes structural and functional information of genes. In Yeast proteins, newly generated domains cover 49.84% more by adjusting e-values and 77.06% more by utilizing GO. In addition, an average number of proteins which share same patterns are reduced to 1.27 with e-value adjusting, 1.55 with utilizing GO. To reduce an average number of GO for one protein, an abstraction rule for GO acyclic graph is also proposed in this paper. Taking all results we have done into consideration, we found that both approaches improve determinability and utilization of proteins, and they can be used complementary.

Keywords: protein interaction, e-value, interpro, gene ontology, abstraction

1 Introduction

With a large amount of biological data published through the Internet, many computational protein-protein interaction prediction methods have been developed. A domain based approach is one of the approaches, and it uses structural information of protein which is known as a domain [3]. Thus, most of domain based approaches infer protein-protein interaction according to domain patterns in protein pairs [1, 2, 4]. Typically, domain based prediction methods gather protein interaction data from several databases [12, 13], and find related domains from other sides [5, 10]. However, a lot of protein interaction information would be abandoned during this process because domains of some proteins have not been discovered yet. In addition, even though two proteins have different functions, they are considered as same unless their domain information is not identical.

Among all possible solutions, one easy but time consuming task is increase domains by a direct extraction of domains from protein raw sequences. InterProScan is a tool for analyzing protein sequences that combines different protein signature recognition methods into one resource. Current InterPro [10] (http://www.ebi.ac.uk/interpro/) frequently publishes domain information screened by a certain expect value. In this paper, we adjusted an expect value of InterProScan and extracted more domains of Yeast proteins. On the other hands, characteristics of protein can be expressed by other alternations such as Gene Ontology (GO) [7]. GO consortium provides structural description of protein functions with diverse views of protein structure, cellular component, molecular function, and biological process, so it may allow more determinable information. GO is annotated on a rooted directed acyclic graph (DAG), and there are hierarchical relationship among the GO terms. These hierarchical relationships mean that an ancestor of a certain node always has a general meaning of the node, and all nodes exist on the same path from the root to leaf may be defined by time-sequential efforts. To validate a feasibility of using GO for a protein prediction, test result was compared to a conventional domain based approach. Since one protein usually has far more GO terms than currently reported domain annotations we pruned away the DAG until average number of GO for one protein becomes small enough to handle in a statistical method. Every leaf node which has a few gene products was abstracted according to a path of GO graph. This abstraction was efficient for reducing an average number of GO terms for one protein from 6.5 to 4.6. As a result of evaluation, In Yeast proteins, newly generated domains cover 49.84% more by adjusting e-values and 77.06% more by utilizing GO. Moreover, a number of proteins sharing same patterns is reduced to 1.27 with e-value adjusting, 1.55 with utilizing GO.

This paper is organized as follows. In Related Work Section, we briefly enumerated other researches to clarify the problem. In Section Method, biological databases we used and validation procedures are illustrated. An abstraction algorithm of GO is also introduced in this section. Our test results is reported in Section Results with a comparison of each idea. Some issues about different approaches are discussed in Section Discussion, and finally we draw conclusion in Section Conclusion.

2 Related Work

Domain based prediction methods and service systems usually integrate protein and domain information into local repository to predict interaction of unknown protein pairs [2]. PreSPI, for an instance, integrates protein interaction data from DIP [12] and domains from [10]. Since this system uses domains to infer protein interaction, PreSPI maps proteins to corresponding domains. One problem during this process is that since this system uses domain information only, two different proteins may be considered as identical. As an example, four different Yeast proteins P25611, P40971, P19541, and P46954 have "IPR001138" domain, so they are considered as same protein in this system. However, UniProt [13] categorizes these proteins into transcriptional regulatory proteins for P25611 and P19541, activator for P19541, and an interaction protein with the SNF1 protein kinase for P46954. "IPR001138" is an accession ID of InterPro which stands for a fungi transcriptional regulatory domain. This domain annotation for a protein seems to be roughly right, however, it is not enough to address slight differences among the proteins. It is natural that domain information cannot be differentiated though an accession ID is changed to other databases because all annotations were made by cross reference mechanisms. To make matters worse, only about 50% protein interaction pairs are usable out of 15,000 pairs in Yeast proteins. This information loss occures, for some domains have not been discovered or relationship between proteins and domains has not been established yet.

Table 1: Diverse domain accession IDs for four proteins

	InterPro	SMART	Pfam	PROSITE
P25611	IPR001138; Fungi_Trscrp_N.	SM00066; GAL4; 1.	PF00172; Zn_clus; 1.	PS00463; ZN2_CY6_FUNGAL_1; 1. PS50048; ZN2_CY6_FUNGAL_2; 1.
P40971	IPR001138; Fungi_Trscrp_N.	SM00066; GAL4; 1.	PF00172; Zn_clus; 1.	PS00463; ZN2_CY6_FUNGAL_1; 1. PS50048; ZN2_CY6_FUNGAL_2; 1.
P19541	IPR001138; Fungi_Trscrp_N.	SM00066; GAL4; 1.	PF00172; Zn_clus; 1.	PS00463; ZN2_CY6_FUNGAL_1; 1. PS50048; ZN2_CY6_FUNGAL_2; 1.
P46954	IPR001138; Fungi_Trscrp_N.	SM00066; GAL4; 1.	PF00172; Zn_clus; 1.	PS00463; ZN2_CY6_FUNGAL_1; 1. PS50048; ZN2_CY6_FUNGAL_2; 1.

As a Table 1 shows, different IDs cannot give more determinability of protein but only support wider choices to use. As a result of that, accuracy of prediction based on domain approach cannot help decreasing. In this paper, we propose two ideas to clear away these obstacles; find more domains which were filtered during extraction from a protein raw sequence and use gene ontology annotations instead of domains. Currently, InterPro [6] screens out some domains with a certain cut-off value to maintain high quality of data. However, it would be better for machine learning systems like a PreSPI to give more training data in spite of having some noises if the noises do not harm prediction accuracy seriously. GO consortium provides functional annotations of proteins as a rooted directed acyclic graph (DAG) form [7]. These annotations are categorized by cellular component, molecular function, and biological process. Each category has GO terms which have hierarchical relationship. As of July 3, 2006, GO reports 10728 biological processes, 1746 cellular components, and 7432 molecular functions. Recently, Minhua Deng, and his colleague used GO annotation in a Markov Random Field (MRF) method to the prediction of Yeast proteins [8]. In their research, a functional path for a GO node was defined as the path from the root to the node, and three categories of GO annotation were used MRF method separately. In a recent approach of Xiaomei Wu et al, they suggested to making a Yeast protein-protein interaction network using GO annotation [9]. In Wus research, they devised a process that defines a similarity between two GO terms in terms of a path distance. Since GO is a DAG, we should define a method which can determine similarity between two GO terms. Note that there are lots of duplicated GO annotations for one protein none the less because only the most specific term should exist. Major method to remove such duplication was finding very informative GO terms. If there are two or more GO terms for one protein on the same path from root to a certain GO term, we selected a leaf node among them. The selected leaf node was evaluated in terms of a count of gene products corresponded to the node. Though a leaf node was found, we replaced the node to an ancestor GO term if a count of gene product is less than 50. Detail validation procedures will be illusterated in Section Methods. Average usability of protein interaction pairs after this abstraction is compared to a conventional domain approach in Section Result. In Section Result, we also report a determinability of GO about a protein is compared to one of domain concept.

3 Methods

In this section, we explain several biological databases which were mainly used in our research. Then, two different validation procedures are described. Both approaches are adopted to increase protein utilization and improve protein determinability.

3.1 Biological Databases

Most of the statistical protein interaction prediction methods need protein interaction data and relevant feature information of proteins such as a domain. In our research, we used such information extracted from several biological databases. Protein interaction data was extracted from the Database of Interacting Proteins (DIP) [12], and domain data was gathered from Integrated documentation resource of Protein families, domains and functional sites (InterPro) [10] and Protein Information Resource (PIR) [11]. To compare a gene ontology concept to a domain based approach, gene ontology information was gathered from iProClass of PIR. iProClass provides diverse accession IDs and related GO terms. Practically, GO consortium only provides Yeast proteins as SGD accession ID, we should find another database to map UniProt accession ID of DIP proteins to GO terms. iProClass is suitable database to easily extract GO terms with UniProt accession ID. Even though it is true that UniProt [13] itself gives us GO terms, it takes a lot of time to store UniProt database locally, so we utilized iProClass instead. DIP is an ideal starting point to extract protein interactions because we can get interaction data of seven representative species *D. melanogaster, S. cerevisiae, C. elegans, H .pylori, H. sapiens, E. coli,* and *M. musculus* and each data has UniProt accession ID which helps us find domain information easily. Besides, they are periodically updated in reasonable span of time (The full sets: a month, CORE subsets: 3 months). In this paper, we use protein interaction data of DIP database released on April 2, 2006, and iProClass database of July 25, 2006. InterPro was initially built in order to integrate scattered databases into one site. Nowadays, its abundant data leads to make itself possible to be a major domain repository. Each organism listed in DIP has a taxonomy ID, and "IPRXXXXXX" formatted domain information can be searched with the ID. The latest release of InterPro contains 12,953 entries and covers 78.1% of UniProtKB which means that 2207141 of 2826393 proteins (as of release 13.0). One way to map DIP ID to InterPro domain ID is using UniProt accession ID which is commonly used for both databases. Since most statistical prediction methods produce good results only when enough features of protein are provided, finding associated domains and gene ontology of a protein is really important. Thus, one of the purposes of this research is to gather plenty of domains and gene ontology information of various species as much as possible.

3.2 InterProScan and e-value adjustment

InterProScan is a tool for analyzing protein sequences that combines different protein signature recognition methods into one resource. It takes sequence data in a recognized sequence format such as raw, FASTA or EMBL as an input and calculates checksum to identify previously analyzed data. If there is no preexisting result matched with checksum, it starts searching using search method such as HMMer and Blast from each of member databases. It figures out information of input sequence and show matches which have similarity with input sequence. InterProScan has member databases such as PROSITE, PRINTs, PFAM, ProDom, SMART, TIGRFAMs, PIR SuperFamily (PIRSF), SUPERFAMILY, Gene3D, and PANTHER. InterProScan shows filtered results which are considered as relevant matches. E-values are used as cut-off values of filtering process in most of databases. As shown in Table 2 the e-values may be different according to each of member databases and scanning methods.

Table 2: Default e-values of member databases

Database	E-value
Pfam	1000
PRINTs	0.001
Gene3D	59.5
Panther	0.001
ProDom	0.01
TIGRFAMs	20
SUPERFAMILY	0.02
SMART	0.01

E-value (expect value) is used as a criterion of significance of matches that is found from subject sequence databases as a result of scanning from the inputs. Generally, it shows a tendency that e-value of sequence having biological meaning is very smaller than 1. It means small number of e-value has high possibility to be a correct hit. E-value tends to be high if database of subject sequence is big, length of input sequence is long, and bit score is small. The reason that the databases have different e-values is details of processing step are different for every database. To take Pfam as an example, it has three processing steps using three cut-off values to find the maximum number of true positive hits during a search but no false positive hits. In first step, e-value is set artificially high because Pfam want to make sure that no true positive hits are missed. As a result of a high e-value, lots of false positives are represented on the result. Second cut-off value is bit score (so called GA) which is manually adjusted by Pfam curators. In this step if bit score is smaller than GA cut-off then it will be removed. After that, if there are Pfam models overlap on the sequence, one of them is removed in a last process called Clan filtering [5, 6]. In this paper, we manually changed e-value of databases in order to extract more hits from raw sequences.

3.3 Gene Ontology and Abstraction

To use gene ontology (GO) terms for a feature of protein, we should be able to extract GO terms as many as possible. Moreover, among the GO terms, there should be some patterns to discriminate between interaction proteins GOs and non interaction proteins' ones. The former requirement of GO was easily convinced by a small test. For seven different species of DIP, we made a statistical table about gene ontology utilization.

Table 3: Statistics of domains and gene ontology terms of interacting proteins in seven different species

	I	II	III	IV	V	VI	VII	VIII	IX	X
Yeast	4953	2187	3897	44.32	78.97	2.05	2.40	2.01	2.09	6.50
Human	872	462	625	52.98	71.67	3.04	3.94	1.88	3.39	9.20
Mouse	197	101	122	51.27	61.93	3.38	3.97	2.03	3.29	9.29
Fruit Fly	1519	487	812	32.06	53.46	2.49	3.78	1.53	2.23	7.54
C.elegans	2141	921	899	43.02	41.99	2.06	0.90	0.62	1.76	3.28
H.pylori	704	133	203	18.89	28.84	2.71	1.81	0.69	3.01	5.52
E.coli	1631	726	824	44.51	50.52	2.36	1.61	0.79	2.67	5.08

I: A count of proteins extracted from interaction protein pairs
II: A count of proteins whose domains are already known
III: A countof proteins whose gene ontology terms are already known
IV: II / I x 100.00
V: III / I x 100.00
VI: An average count of domains for one protein
VII: An average count of molecular function in gene ontology for one protein
VIII: An average count of biological process in gene ontology for one protein
IX: An average count of cellular component in gene ontology for one protein
X: An average count of all categories in gene ontology terms for one protein

As shown in Table 3, interacting proteins have sufficient GO terms. However, the amount of GO terms is too large to construct a learning set. In yeast proteins, an average count GO for one protein is over 6.0 (in the column X). In some prediction method, especially in PreSPI, this high average count may lead to a huge space complexity. One alternative way is that we adapt only one category of gene ontology. Nevertheless, we decided to use all categories in order to merge structural and functional features of proteins. Since GO terms have been annotated by different manners for a while, one proteins may have several GOs existing on the same path from root to a leaf. Note that GO graph has hierarchical relation between ancestors and children. In this case, we need a treatment about selecting the most representative GO. Thus, following abstraction rules were devised.

1. Remove unknown gene ontology terms
2. Remove all ancestors which exist in the same path from root to the gene ontology
3. Replace leaf nodes to their ancestors until they have more than 50 gene products
4. Dont abstract when the gene ontology term is direct child of category gene ontology term

In the gene ontology, each category has unknown terms. Biological process unknown is GO:0000004, molecular function unknown is GO:0005554, and cellular component unknown is GO:0008372. Since

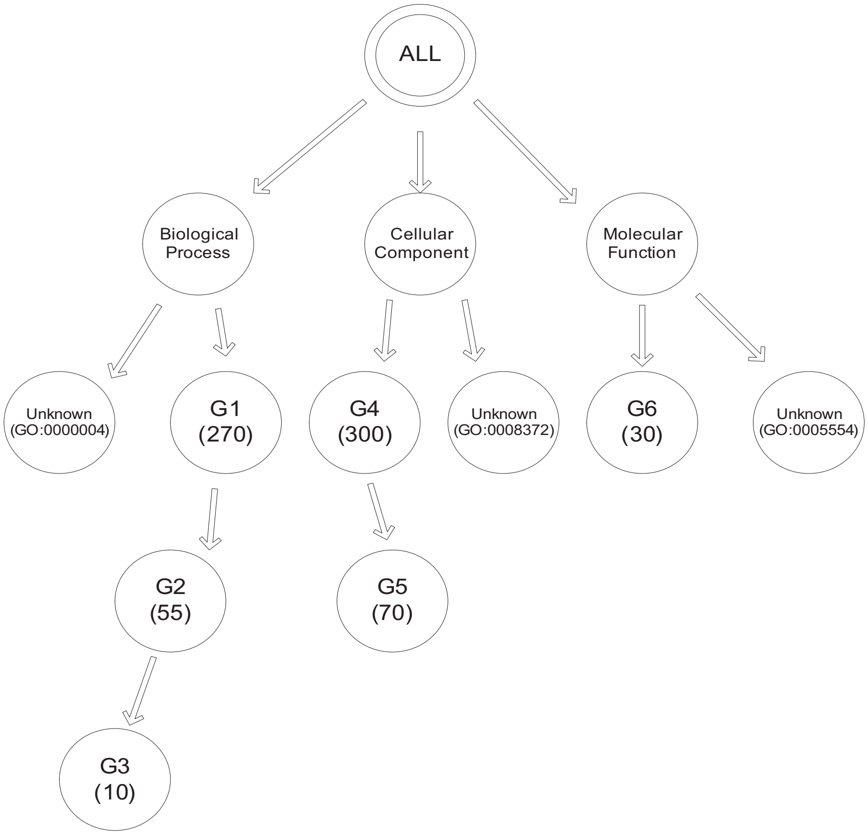

Figure 1: An example graph of GO ($G1 - G6$: GO terms and a count of their gene product).

these terms are not informative at all, we eliminated them firstly. In addition, if a protein has two or more GOs for a one category, we should inspect a correlation among the GOs. Practically, some proteins have unique ontology among the same path from the root to the leaf, but some proteins have two or more ontology terms which exist on the same path from the root to the leaf node. In this situation, if one term is a child of another term, we can remove a parent node because child node has more specific meaning. After that, we reconfigured GO graph to have informative nodes only. In our research, when a node has more than 50 gene products, we considered the node as informative. 4152 children among 13763 were replaced to their parents due to this rule. A cut off value 50 of gene products is remained here as an open question. In the future, we can adjust the value to make a learning set has reasonable size. This abstraction is applicable only when a GO node is not a direct child of category GO term such as "GO:0008150", "GO:0005575", or "GO:0003674" which stand for biological process, cellular component, and molecular function, respectively. Suppose that there are two proteins, A and B, and related GO terms are $A = \{G1, G3, G4, G6\}$ and $B = \{G3, G4, G5\}$. If a gene ontology graph looks like a Figure 1, GO terms of protein A were firstly change to $\{G3, G4, G6\}$ due to rule 2. Then, by the rule 3, $G3$ is replaced to its parent $G2$. Note that $G6$ is not changed during this process because of rule 4. Likewise, GO terms of protein B is changed to $\{G2, G5\}$. In Section Result, we represent an abstraction result of interaction proteins of Yeast.

4 Result

In this section, we enumerate comparisons between a domain and gene ontology approaches. As we can see in Table 4, both e-value adjusting and GO replacing increase a utilization of proteins in Yeast.

Table 4: Overall comparison between a domain and gene ontology approaches

	Maximum count of domains or GO terms	Count of domains or GO terms	Count of usable proteins	Average proteins sharing same patterns	Average domains or GO per one protein
Original	11	1123	2201	1.96	2.05
eVal	13	2588	3298	1.27	2.36
Normal GO	22	2514	3897	1.55	6.50
Abstraction	18	1218	3895	3.20	4.60

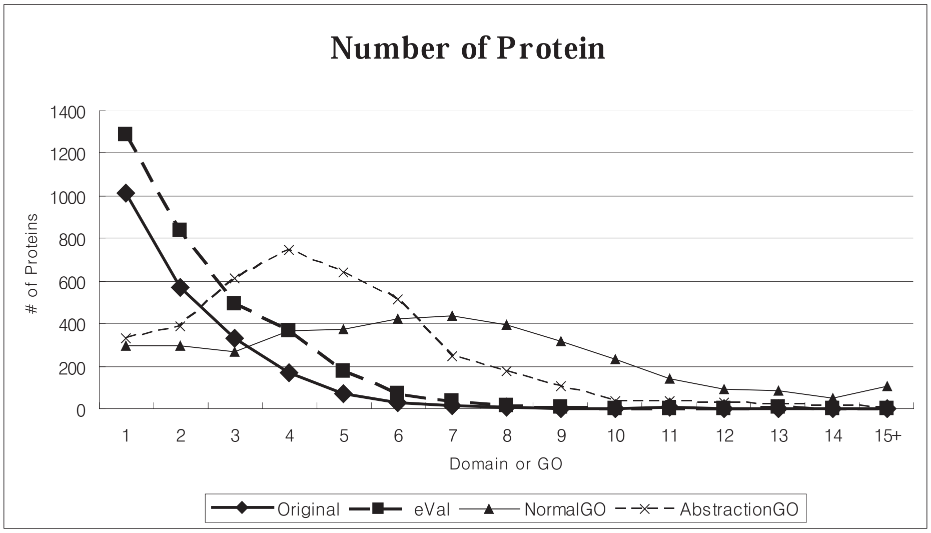

Figure 2: A count of proteins which have X domains or GO terms

Adjusting improves a utilization of proteins by 49.84% , and GO also enhances by 77.06%. Moreover, proteins sharing same patterns were commonly reduced in both methods. Abstraction rule we devised was really good in the point that it decreasesd average number of GOs per one protein without any harm about protein utilization. However, it may reduce diversity of proteins because the abstraction result shows that more proteins share same GO patterns. (1.55 to 3.20) Table 5 and Figure 2 show a comparison of a count of proteins. In a conventional domain based approach, most of proteins have a small number of domains while proteins are well distributed to diverse number of GO terms in a new approach. As a result of changing cut of value of InterPro machine, a total count of usable protein was enormously increased. From an abstraction row in Table 6, we found that even though abstraction algorithm decrease an average count of GO terms for one protein, it may harm to a diversity of interaction proteins.

Table 5: A count of proteins which have X domains or gene ontology terms

X	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15+
Original	1010	569	334	171	68	25	11	7	1	1	4	0	0	0	0
eVal	1286	835	492	368	175	73	36	16	8	3	1	1	4	0	0
Normal GO	297	298	270	369	374	425	439	395	317	233	143	94	85	51	107
Abstraction	329	390	614	746	641	513	248	178	103	32	33	28	19	13	8

Table 6: A count of domains or gene ontology patterns in the proteins which have X domains or GO terms

X	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15+
Original	400	313	179	104	43	22	8	6	1	1	2	0	0	0	0
eVal	811	528	334	220	125	56	33	13	6	3	1	1	2	0	0
Normal GO	51	167	231	332	327	366	401	365	299	210	136	90	76	47	95
Abstraction	55	230	458	619	577	457	236	171	92	31	33	22	11	12	8

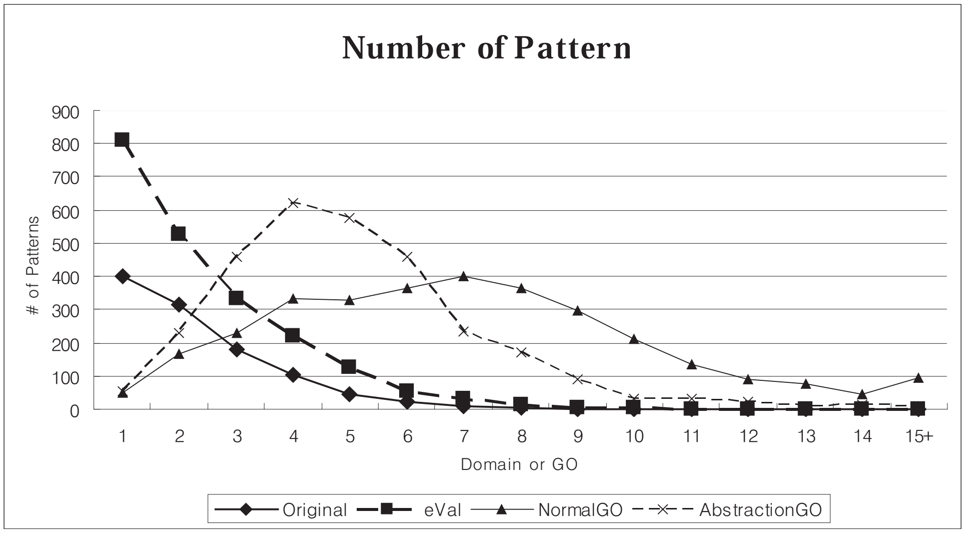

Figure 3: A count of domains or gene ontology patterns in the proteins which have X domains or GO terms

5 Discussions

Protein interaction prediction methods become sensitive if they can consider more diverse features of protein. In our research, both a changing e-value of InterProScan and a using GO terms without abstraction showed the best results in the sense of diversity of feature patterns. However, an amount of usable protein is much larger in GO adaptation so it can give us large learning set. In this paper, we used all categories of GO; Biological Process (BP), Cellular Component (CC), and Molecular Function (MF). In comparison to conventional domain approach, more diverse information about protein can be considered, but we should observe differences when we use each category independently for a comprehensive study. Obtaining hidden domains using adjusting e-value have several limitations. One of them is that it is hard to adjust proper line of e-values because all of member databases have their own e-values and they have post-processing steps their own. This diversity of member databases makes difficulties to adjust proper line of e-values. Another one is that annotation process of InterPro is doing manually. This manual step means that proper InterPro accession ID cannot be annotated immediately even if we find a reference ID of other biological databases. In other words, there may be no InterPro accession ID correspond to domains that we found with low e-value. Note that low e-value has high probability to possess a biological meaning. Table 7 shows an example of analysis sequence, hidden domains resulted from filtering process. Protein P09798 reported that it has four domains, IPR001440, IPR008940, IPR011990, and IPR013026 in InterPro. Table 7 shows other domains which can be found in different e-values. PF07719 of Pfam database can be annotated by InterPro ID, IPR013105 with a reasonable e-value. If we make e-value very high, PD000191 of ProDom can be found as well. However, subsequences of Panther, SUPERFAMILY, and Gene3D cannot be found from InterProScan because they have no correspond InterPro ID in spite of very high realiability.

Table 7: A result of analysis protein, P09798

Database	ID of Database	E-value	InterPro ID
Panther	PTHR12558	6.8e-123	NULL
SUPERFAMILY	SSF48452	3.7e-43	NULL
Gene3D	1.25.40.10	1.9e-42	NULL
Pfam	PF07719	1.1e-05	IPR013105
ProDom	PD000191	1e+01	IPR001963

6 Conclusion

Most of machine learning approaches should have enough size of learning sets to predict protein-protein interaction. Similarally, domain based prediction methods should obtain sufficient protein-protein interaction data and relevent domains as well. However, since domains are not well discovered yet, we suffered from loss of protein interaction data. In conventional domain approaches, proteins are categorized by their domain patterns, so we cannot distinguish two different proteins if they have same domain patterns. In this paper, we proposed two different ideas; adjusting e-value of InterProScan and using GO instead of domains. Both approaches are good for increasing utilization of proteins, and they improved diversity of proteins. We also divised an abstraction algorithm of GO graph, and it reduce a space complexity of GO terms. Currently, the result of an abstraction algorithm doesn't show very good result we expected in the sense of improving diversity of proteins. However, we convinced that the abstraction rule decrease an average number of proteins per one protein without any harm about a protein utilization. In the future, we will evaluate our approaches to practical prediction system and report accuracy differences.

References

[1] M. Deng, S. Metah, F. Sun and T. Chen, Inferring Domain-Domain Interactions from Protein-Protein Interactions, *Genome Research*, 12:1540–1548, 2002.

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

[3] E. Sprinzak and H. Margalit, Correlated sequence-signatures as markers of protein-protein interaction, *J. Mol. Biol.*, 311:681–692, 2001.

[4] S. Ng, Z. Zhang and S. Tan, Integrative approach for computationally inferring protein domain interactions, *Bioinformatics*, 19:923–929, 2003.

[5] Robert D. Finn, et al., Pfam: clans, web tools and services, *Nucleic Acids Research*, 34:D247–D251, 2006.

[6] Quevillon E., et al., InterProScan: protein domains identifier, *Nucleic Acids Research*, 33:W116–W120, 2005.

[7] Ashburner M., et al., Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nature Genet*, 25:25–29, 2000.

[8] Minghua Deng, et al., Mapping gene ontology to proteins based on protein-protein interaction data, *Bioinformatics*, 20:6, 2004.

[9] Xiaomei Wu, et al., Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations, *Nucleic Acids Research*, 34:7, 2006.

[10] R. Apweiler, et al., The InterPro database, an integrated documentation resource for protein families, domains and functional sites, *Nucleic Acids Research*, 29:37–40, 2001.

[11] Cathy H. Wu, et al., The Protein Information Resource, *Nucleic Acids Research*, 31(1):345–347, 2003.

[12] I. Xenarios, et al., DIP: The Database of Interacting Proteins: 2001 update, *Nucleic Acids Research*, 29:239–241., 2001.

[13] Bairoch A., et al., The Universal Protein Resource (UniProt), *Nucleic Acids Research*, 33:D154–D159, 2005.