Journal of Bioinformatics and Computational Biology
© Imperial College Press

AN INTEGRATIVE DOMAIN-BASED APPROACH TO PREDICTING PROTEIN-PROTEIN INTERACTIONS

Thanh Phuong Nguyen and Tu Bao Ho

Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923-1292 Japan
{*phuong,bao*}*@jaist.ac.jp*

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

Protein-protein interactions (PPI) are intrinsic to almost all cellular processes. Different computational methods offer new chances to study PPI. To predict PPI, while the integrative methods use multiple data sources instead of a single source, the domain-based methods often use only protein domains. Integrating both protein domain features and genomic/proteomic features from multiple databases could be more powerful in PPI prediction. Moreover, it can allow discovering reciprocal relationships between PPI and biological features of their interacting partners.

We develop a novel integrative domain-based method for predicting protein-protein interactions using inductive logic programming (ILP). Two principal domain features used are domain fusions and domain-domain interactions. Various relevant features of proteins are exploited from five popular genomic and proteomic databases. By integrating these features, we constructed biologically significant ILP background knowledge of more than 278,000 ground facts. The experimental results through multiple 10-fold cross-validation demonstrated that our method can better predict protein-protein interactions than other computational methods in terms of typical performance measures. The proposed ILP framework can be applied to predict domain-domain interactions with high sensitivity and specificity. The induced ILP rules give us a lot of interesting biological reciprocal relationships among PPI, protein domains, and genomic/proteomic features related to PPI.

Supplementary materials: `http://www.jaist.ac.jp/%7Es0560205/PPIandDDI/`

Keywords: protein-protein interaction; domain-domain interaction; inductive logic programming.

1. Introduction

Protein-protein interactions are indispensable at almost every level of cell function, in the structure of sub-cellular organelles, in the transport across the various biological membranes, in muscle contraction, signal transduction, and regulation of gene expression, etc. Detecting protein functions via prediction of protein-protein interactions (PPI) has emerged as a new trend, both *in vitro* and *in silico*. Therefore, prediction of protein-protein interactions has become one of the most challenging tasks in the post-genomic era. Experimental techniques have marked unmistakable progress in finding out and verifying protein interactions for diverse organisms, including well-known ones such as two-hybrid assay,[1,2] affinity purification and mass spectrometry,[3] phage display.[4] Because of little overlap among these experimental databases, the question about their reliability is raised.

With the recent blooming of public proteomic and genomic databases, numerous computational approaches offer a chance to study more widely and deeply regarding protein-protein interactions. Depending on the source of information used, computational approaches can be categorized in three groups: structure-based approach, sequence-based approach, and genome-based approach, typically the work of,[5,6,7] respectively. Besides methods based on a single data source, many bioinformaticians make the effort in the *integrative approach* that employs multiple data sources to better predict PPI. Jansen *et al.* used a Bayesian network approach for integrating weakly predictive genomic features into predictions of protein-protein interactions.[8] Several kernels for different data sources like protein sequences, Gene Ontology annotations, local properties of networks, etc. are combined to infer PPI.[9] Some other efforts were probabilistic decision tree approach,[10] inductive logic programming method,[11] probabilistic model,[12] etc. From multiple data sources, these works can extract and combine various genomic and proteomic features related to PPI. The obtained results showed many advantages of multiple data source integration.

Protein domains are structural and/or functional units of proteins that are conserved through evolution to represent protein structures or functions. They are the key regulators in protein-protein interactions. Interactions among domains are needed as stable channels of PPI. Recently, the *domain-based approach* to prediction of PPI has received much attention in many ongoing studies. One of the pioneering works based on protein domains is an association method.[13] Kim *et al.* improved the association method by considering the number of domains in each protein.[14] Han *et al.* proposed a domain combination-based method by considering the possibility of domain combinations appearing in both interacting and non-interacting sets of protein pairs.[15] A graph-oriented method is proposed by Wojcik and Schachter called the interacting domain profile pairs (IDPP) method.[16] Chen *et al.* used domain-based random forest framework to predict PPI.[17] Martin *et al.* used signatures generated from sequences to predict PPI, and they constructed domain-sized amino acid subsequences through sliding a window across each of the protein sequences to predict domains.[18]

The shortcoming of integrative methods is that they do not take protein domains into account while there are evidences that the biological mechanism underlying protein-protein interactions involves protein domains and their interactions.[19] On the other hand, while domain-based methods all treasured the biological roles of protein domains in PPI prediction, most of them merely considered the co-occurrence of domains/domain pairs. To comprehensively predict PPI it seems necessary that domain-based methods could also employ genomic/proteomic features.

This work, early initialized in,[20] presents a novel integrative domain-based approach using inductive logic programming to predict protein-protein interactions. The key idea of this computational method is to integrate protein domain features and multiple genomic/proteomic features. To efficiently integrate such two kinds of feature in predicting PPI, we specified two main tasks. The first is to extract as many as possible useful domain and genomic/proteomic features related to PPI. From seven popular databases, we extracted more than 278,000 ground facts of domain fusion, domain-domain interaction features and various other biologically significant genomic/proteomic features. The second is to employ inductive logic programming (ILP) with the huge amount of background knowledge to effectively infer PPI.

To demonstrate the advantages of the integration domain features and genomic/proteomic features in PPI prediction, we conducted multiple 10-fold cross validation for comparing our methods with two other methods based on single domain features, as well as with the non domain-based approach using multiple genomic databases. The performance measures include ROC curves, sensitivity and specificity. In all cases, our method performed considerably better than the others. Furthermore, with numerous protein and domain data, domain-domain interactions were successfully inferred by our method with high sensitivity and specificity. At last, analyzing various produced rules (of both PPI and DDI), many interesting relationships among PPI and DDI and protein functions, biological processes, conserved motifs and pattern sites were found. Our proposed methods can be tuned to predict PPI and DDI for diverse organisms and other genomic and proteomic data sources.

The remainder of the paper is organized as follows. In Section 2, we present our proposed method to predict PPI based on domains using ILP and multiple genomic and proteomic databases. The comparative evaluation of the experiments is given in Section 3. Predictive rules of PPI and DDI, as well as discussion, are presented in Section 4. Some concluding remarks are given in Section 5.

2. Materials and Methods

In this section, we present our proposed method to predict protein-protein interactions based on domain and multiple genomic/proteomic data using ILP. Two main tasks of the method are: (1) Constructing integrated background knowledge[a] of domain features and multiple genomic/proteomic features, and (2) Learning PPI predictive rules by ILP from the constructed background knowledge. Constructing ILP background knowledge requires two steps. The first one is defining ILP predicates. The second one is extracting ground facts to extensionally define predicates. When choosing a feature, we concentrated on two points: (i) the biological role of that feature in protein-protein interactions or domain-domain interactions,

[a] the terms ‘background knowledge’ and ‘ground facts’ (the second task) are used in terms of the language of inductive logic programming.

and (ii) the availability of data for that feature. Consulting results of experimental and computational research on PPI, twenty two features of protein domains and genomes/proteomes are chosen and are formulated using ILP predicates. The large database of more than 278,000 ground facts of twenty two predicates is sufficient for accurate PPI prediction.

We first briefly introduce about Inductive Logic Programming and some bioinformatics applications of ILP in Section 2.1. Then the first task in our proposed method is presented in Subsections 2.2, 2.3, and 2.4. Subsection 2.5 describes the second task.

2.1. *Inductive Logic Programming*

Inductive logic programming is the intersection of machine learning and logic programming.[21] Inductive logic programming uses logic programming as a uniform representation for examples, background knowledge and hypotheses. Given an encoding of the known background knowledge and a set of examples (positive and negative examples) represented as a logical database of ground facts, an ILP system will derive hypotheses in form of logical rules which entails all the positive and none of the negative examples. The schema of ILP as following

$$\textit{Positive examples} + \textit{Negative examples} + \textit{Background knowledge} \Rightarrow \textit{Hypotheses}$$

Distinguishing features of ILP are its ability to take into account background (domain) knowledge in the form of logic programs, and the expressive power of the language of discovered patterns.[22]

There have been many ILP systems that were applied to different problems in bioinformatics. ILP is particular suitable for bioinformatics tasks because of its ability to take into account background knowledge and work directly with structured data.[23] The ILP system GOLEM has been applied to find the predictive theory about the relationship between chemical structure and activity.[24] Other central concerns of bioinformatics have been convincingly solved by ILP, such as protein secondary structure prediction,[25] protein fold recognition,[26], etc. With abundant biological and biomedical data, ILP is potential to appropriately combine them in background knowledge to settle various problems in bioinformatics.

2.2. *Extracting domain fusion and domain-domain interaction data*

Protein domains form the structural or functional units of proteins that partake in intermolecular interactions. The existence of certain domains in proteins can therefore suggest the propensity for the proteins to interact or form a stable complex to bring about certain biological functions. Because of their important biological roles in PPI prediction,[19,27] domain fusion and domain-domain interaction features are used in our work.

Let P denote the set of considered proteins p_i. Denote by D the set of all protein domains d_k which belong to proteins p_i. A pair of interacting proteins (p_i, p_j) is denoted by p_{ij}, and a protein pair that does not interact with each other by $\neg p_{ij}$.

Domains of interacting proteins have more chance to fuse together than domains of non-interacting proteins. Therefore, when finding a pair of proteins which have fused domains, we can predict an interaction between them.[28] Domain fusion data is referred from Domain Fusion Database.[29] Truong *et al.* employed relational algebra to find domain fusions in protein sequence databases. We extracted domain fusion data for all protein pairs (p_i, p_j), $p_i, p_j \in P$. The following predicate represents the domain fusion between two proteins

```
domain_fusion(+protein, +protein, #FUSION)
```

Note that in the ILP system used – the learning engine Aleph for proposing hypothesis[30] – there are some *mode declarations* to build the bottom clauses, and a simple mode type is one of the following: (1) *the input variable* (+), (2) *the output variable* (−), or (3) *the constant term* (#). Predicate `domain_fusion` means whether two input proteins, A and B, have fused domains or not (valued "yes" by the constant term `#FUSION`). This predicate is supported by a set of ground facts G_{domain_fusion}, e.g., domain_fusion (ap3m_yeast, ap3b_yeast, yes). After preprocessing, the set G_{domain_fusion} consists of 2,761 ground facts.

Let d_{kl} and $\neg d_{kl}$ denote a domain-domain interaction and a non-interacting of a domain pair (d_k, d_l), respectively. The assumption that proteins interact with each other through interactions of their domains is widely accepted and validated. The domain-domain interaction data is exploited to more reliably predict PPI. We extracted DDI data from **iPfam** database (http://www.sanger.ac.uk/Software/Pfam/iPfam/) that is a resource describing domain-domain interactions observed in PDB entries. The domains are defined by Pfam. When two or more domains occur in a single structure, the domains are analysed to see if they form an interaction. If the domains are close enough to form an interaction, the bonds forming the interaction are calculated.

We considered two features of DDI. The first feature is whether a protein pair (p_i, p_j) has a domain interaction d_{kl}, and if yes, how many d_{kl} it has. This information is formulated by predicate

```
hasddi(+protein, +protein, #DDI)
```

where the `#DDI` value is the number of DDI mediating the same PPI p_{ij}. The set of ground facts for this predicate G_{ddi} includes 657 ground facts, some of them are: hasddi(jsn1_yeast,yip1_yeast,2), hasddi(msh4_yeast,msh5_yeast,5), etc.

The number of domain-domain interactions of a protein is one of the features which may increase or decrease the probability of its interaction with others. So we considered the relationship between PPI and the number of DDI of each interacting partner. This relationship is presented in the following predicate

```
num_ddi(+protein, #NUM_DDI)
```

Denoted by G_{num_ddi} the set of ground facts of the above predicate contains 505 ground facts. We found that there are some proteins having a large number of DDI, for example, num_ddi(did4_yeast,20) or num_ddi(bud27_yeast,39), and these proteins potentially interact with many other proteins.

2.3. *Extracting proteomic and genomic data from multiple databases*

In addition to domain fusion and domain-domain interaction features, we mined genomic and proteomic data from UniProt database, CYGD database, InterPro database, Gene Ontology database, and Gene Expression database to detect useful genomic and proteomic features for PPI prediction. Table 1 shows 19 predicates corresponding to genomic/proteomic data extracted from multiple databases.

As the world's most comprehensive catalog of information on proteins, **UniProt database** (http://www.pir.uniprot.org/) largely provides functional, structural or other categories (in Keyword - KW line); regions or sites of interest in the sequences (in Feature Table - FT lines); describes enzymes coded (EC). Others are pointers to information related to entries and found in data collections other than UniProt such as GO database, PIR database, PROSITE database, Pfam database, and Interpro database (in Database cross-Reference - DR line).

Some examples of extracted data of these predicates are keyword(ace1_yeast, transcription regulation), feature(ldb7_yeast, chain chromatin structure remodeling complex), coded_enzyme(uqcr1_yeast, ec1.10.2), and dr_go(twoa5d_yeast,go0005935), etc. The first three predicates present general protein features that should effect their interactions. The other give references to other databases. Data from different databases related to PPI are bound by these predicates.

The MIPS Comprehensive Yeast Genome Database (**CYGD**) (http://mips.gsf.de /genre/proj/yeast/) presents information on the molecular structure and functional network of the entirely sequenced, well-studied model eukaryote, the budding yeast *Saccharomyces cerevisiae.* Among various information provided by CYGD, the following should be mined to discover the relationship between CYGD's categories and protein-protein interactions, i.e. category of functions, category of subcellular locations, category of phenotypes, category of complexes, and category of proteins. A protein has more chance to interact with proteins in the same category than with proteins in different catagories. Here are some examples: subcell_cat (ahc1 yeast, cytoplasm), phenotype_cat(cyk2 yeast, cell cycle defects), etc.

InterPro database (http://www.ebi.ac.uk/interpro/) is a database of protein families, domains and functional sites. We considered the association between InterPro annotations and GO terms. For example, interpro_go(ipr000009,go0007165), interpro_go(ipr000009,go0000159).

Gene Ontology database (http://www.geneontology.org/) has three organizing principles: molecular function, biological process and cellular component. The terms

Table 1. Predicates used as background knowledge in various genomic/proteomic data sources

Database	Background knowledge predicates	#Ground fact
UniProt	`keyword(+protein,#Keyword)` A protein has a proteins keyword	43,539
	`feature(+protein,#Feature)` A protein has a protein feature	
	`coded_enzyme(+protein,#EC)` A protein has a coded enzyme	
	`dr_prosite(+protein, -PROTSITE_ID)` A protein has a PROSITE annotation number	
	`dr_interpro(+protein, -INTERPRO_ID)` A protein has an InterPro annotation number	
	`dr_go(+protein,-GO_TERM)` A protein has a GO term	
	`dr_pfam(+protein, -PFAM_ID)` A protein has an Pfam annotation number	
	`dr_pir(+protein, -PIR_ID)` A protein has a Pir annotation number	
CYGD	`subcell_cat(+protein, #SUBCELLCAT)` A protein has subcellular structures in which it is found	11,909
	`function_cat(+protein, #FUNCAT)` A protein has a certain function category	
	`protein_cat(+protein, #PROTEINCAT)` A protein has a certain protein category	
	`phenotype_cat(+protein, #FENCAT)` A protein has a certain phenotype category	
	`complex_cat(+protein, #COMPLEXCAT)` A protein has a certain complex category	
InterPro	`interpro_go(+INTERPRO_ID, -GO_TERM)` Relation of InterPro annotations and GO terms	4,965
GO	`is_a(+GO_TERM,-GO_TERM)` `is_a` relation between two GO terms	1,142
	`part_of(+GO_TERM,-GO_TERM)` `part_of` relation between two GO terms	
Gene Expression	`expression(+protein, +protein, #COEFFICIENT)` Gene expression correlation coefficient of two proteins	200,000
DIP	`num_ppi(+protein, +protein, #NUM_PPI)` A protein has a number of protein-protein interactions	13,376
	`ig(+protein, +protein, #IG)` Interaction generality of two proteins is the number of protein that interact with just two considered proteins	

in an ontology are linked by two relationships, *is_a* and *part_of*. The relationships of interacting partners in a PPI may effect their interaction. Some ground facts are is_a (go0000002, go0007005), part_of (go0000032, go0007047)).

Interacting proteins are often co-expressed, and then gene expression coefficients between two proteins are useful in predicting PPI. The **Gene Expression** coefficients between two proteins are referred to Jansen *et al.*'s work[8] which contains 25,000,000 pairwise coefficients for about 18,773,128 protein pairs. In our work, we randomly extracted 200,000 gene expression coefficients in terms of ground facts represented by predicates `expression(+protein, +protein, #COEFFICIENT)` for about 11,000 positives and negatives in the training data set.

Two last predicates represent information about the number of protein-protein interactions and interaction generality of two interacting partners. Interaction generality is the number of proteins that interact with both interacting partners in an interaction. The interacting pairs from DIP core data set (see more in Section 3.1) are extracted corresponding to these predicates.

2.4. *Constructing background knowledge for predicting protein-protein interactions*

After defining 22 predicates, we exploit data in terms of ground facts for these predicates from 7 databases (2 databases for domain features and 5 others for genomic and proteomic features). In succession, we denote the sets of ground facts extracted from UniProt database, CYGD database, InterPro database, Gene Ontology database, and Gene Expression database by $G_{UniProt}$, G_{GO}, $G_{InterPro}$, G_{CYGD}, and $G_{expression}$, respectively. Algorithm 1 presents the procedure to extract data from multiple databases to construct background knowledge for PPI prediction.

Algorithm 1 Extracting domain feature data and genomic /proteomic feature data from multiple sources.

```
Input:
  Set of proteins $\{p_i\} \subseteq P$.
Output:
  Sets of ground facts $G_L = \{G_l\}, G_l \in \{G_{domain\_fusion}, G_{ddi}, G_{num\_ddi}, G_{UniProt}, G_{CYGD}, G_{InterPro}, G_{GO}, G_{expression}, G_{ig}, G_{num\_ppi}\}$.

Initialize all sets of ground facts $G_l := \emptyset$; $D := \emptyset$.
Extract all domains $d_k$ belonging to proteins $p_i$; $D := D \cup \{d_k\}$.
for each protein pair $(p_i,\ p_j)$
  for all $d_k \in p_i$ and $d_l \in p_j$
    if $fused(d_k, d_l) = $ true then
      $G_{domain\_fusion} := G_{domain\_fusion} \cup \{(p_i,\ p_j)\}$.
    if $\exists\ d_{kl}$ then
      $G_{ddi} := G_{ddi} \cup \{(p_i,\ p_j)\}$
      Count the number of DDI for proteins $p_i$ and $p_j$ for $G_{num\_ddi}$, respectively.
for each protein $p_i \in P$
  Extract $G_{UniProt}$ and $G_{CYGD}$ from UniProt and CYGD database, respectively.
  Extract mapping data between GO terms $g_i$ and Interpro identifiers $t_i$ related to $p_i$ from
  InterPro database for $G_{Interpro}$; $G_{InterPro} = G_{InterPro} \cup \{t_i, g_i.\}$.
for each protein $p_i \in P$
  for each protein $p_j \in P$
    Extract the relationship $r_{ij}$ between GO terms $(g_i,\ g_j)$ related to $(p_i,\ p_j)$ from
    GO database; $G_{GO} = G_{GO} \cup \{r_{ij}(g_i, g_j)\}$.
    Extract the expression correlation coefficients $e_{ij}$ of $(p_i, p_j)$;
    $G_{expression} = G_{expression} \cup \{p_i, p_j, e_{ij}\}$.
    Extract the interaction generality of PPI $n_{ij}$ of $(p_i,\ p_j)$; $G_{ig} = G_{ig} \cup \{p_i, p_j, n_{ij}\}$.
    if $\exists\ p_{ij}$ then
      $num\_ppi_i := num\_ppi_i\ +\ 1$;
  $G_{num\_ppi} := G_{num\_ppi} \cup \{(p_i, num\_ppi_i)\}$.
return $G_L$.
```

2.5. *Predicting protein-protein interaction with integrative domain-based ILP framework*

Algorithm 2 describes the integrative domain-based ILP framework for predicting PPI from multiple genomic/proteomic databases.

Algorithm 2 An integrative domain-based ILP framework for PPI prediction

```
Input:
    Set of protein-protein interactions $S_{interact} = \{p_{ij}\}$
    Number of negative examples $(\neg p_{ij})$ $N$
    Sets of ground facts $G_{domain\_fusion}$, $G_{ddi}$, $G_{num\_ddi}$, $G_{UniProt}$, $G_{CYGD}$, $G_{InterPro}$, $G_{GO}$,
    , $G_{expression}$, $G_{ig}$, and $G_{num\_ppi}$.
Output:
    Set of rules $R$ for protein-protein interaction prediction.
1: $R := \emptyset$.
2: Extract positive examples for the set $S_{interact}$.
3: Generate N negative examples $\neg p_{ij}$; $S_{\neg interact} = \{\neg p_{ij}\}$.
4: call Algorithm 1 to generate sets of ground facts $G_l$; $S_{background} = G_L = \{G_l\}$.
5: Run an ILP program with $S_{interact}$, $S_{\neg interact}$ and $S_{background}$ to induce rules $r$.
6:     $R := R \cup \{r\}$.
7: return $R$.
```

After initializing the set of rule R in Step 1, Step 2 and Step 3 are for generating positive and negative examples $S_{interact}$ and $S_{\neg interact}$, respectively (see more in Subsection 3.1). In Step 4, we constructed background knowledge $S_{background}$ including both domain features and genomic/proteomic features from sets of ground facts of defined predicates (see Section 2.4). In Step 5, in our experiments, system Aleph was applied to induce rules.

Aleph is an advanced ILP system that uses a top-down ILP covering algorithm. Aleph requires three input files to construct theories: positive examples, negative examples and background knowledge. Positive and negative examples can simply be considered as ground facts. Background knowledge is in the form of Prolog clauses that encode information relevant to the domain. All predicates appearing in hypothesized clauses have to be declared, and amongst them the target predicate is learned to induce rules.

The target predicate in our work is `has_int(+protein, +protein)`, meaning that two arbitrary proteins interact. Aleph learns three inputs and induces rules (hypothesized clauses) in terms of the relationships between the target predicate and other predicates declared in background knowledge.

3. Evaluation

We concentrate on predicting PPI for *Saccharomyces cerevisiae*, a budding yeast. We carried out experimental comparative evaluation, consisting of two experiments corresponding to protein-protein interaction prediction in Section 3.1 and domain-domain interaction prediction in Section 3.2.

3.1. *Predicting protein-protein interactions*

3.1.1. *Experiment design of protein-protein interaction prediction*

To assess the performance of our method for PPI prediction, we did two comparative tests to demonstrate: (1) the advantages of the integration of multiple proteomic and genomic features in our method, (2) the advantages of domain-based approach. The 10-fold cross validation was produced 10 times with each of two negative sets to compare our proposed method with other domain-based methods, particularly AM method and SVMs method. Second, we conducted 10-fold cross validation tests for an ILP method with multiple genomic databases, but not using domain features,[11] and compared those results with our method in terms of sensitivity and specificity.

In two comparative tests with AM and SVMs method, we used the core data of DIP data set (http://dip.doe-mbi.ucla.edu/). This is a large, reliable set of interactions each of which was observed by at least three different methods. Each interaction in DIP database is originally presented by ORF name (Open Reading Frame). After excluding all interactions in which either bait ORF or prey ORF is not found in UniProt database, the positive set has 5,512 interacting pairs from the original 5,963 pairs. We generate two sets of negatives according to two popular methods.[31] In each set of negatives, 5,512 examples to form a training set of equal numbers of negatives and positives. The first one is selecting randomly 5,512 protein pairs from the protein set P where negative examples $\neg p_{ij}$ not belonging to the positive example set $S_{interact}$. The second one is selecting 5,512 protein pairs (p_i, p_j) where two protein, p_i and p_j, are located in different subcellular compartments. In the test with the negatives generated by the second method, we excluded the predicate `subcell_cat(+protein, #SUBCELLCAT)`. Then, the negative set of the second test were assured to be independent with the background knowledge.

3.1.2. *Result of protein-protein interaction prediction*

With the same positives and negatives data sets, we conducted 10-fold cross validation tests for our method, AM method and SVMs method. AM method calculated the probability of protein pairs based on protein domains.[13] In our experiment, the probability threshold is set to 0.05. For SVMs method, we used SVM^{light}.[32] The linear kernel with default values of the parameters was used. For Aleph, we selected *minpos* = 2 and *noise* = 0, i.e. the lower bound on the number of positive examples to be covered by an acceptable clause is 2, and there are no negative examples allowed to be covered by an acceptable clause. We also used the default evaluation function *coverage* which is defined as $P-N$, where P, N are the number of positive and negative examples covered by the clause, respectively.

The ROC curves of ILP, AM and SVMs methods with 5,512 randomly selected negative examples are shown in Figure 1. ROC curve (Receiver Operating Characteristic curve) shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity). Sensitivity refers to

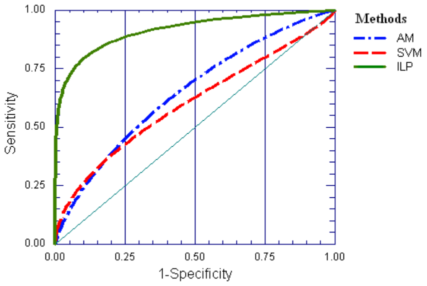

Fig. 1. Comparative ROC curves of ILP, SVMs and AM method with 5,512 random negative examples.

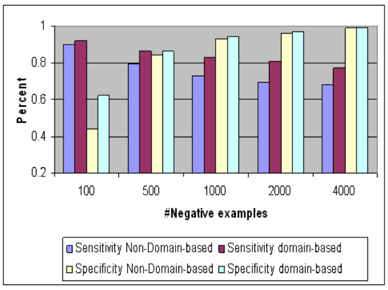

Fig. 2. Comparison of sensitivity and specificity of non-domain based method and our proposed method with various sets of negative examples by 10 times 10-fold cross-validation.

the ability of the test to detect individuals who actually have the disorder. On the other hand, the term specificity means that the test is specific to the disorder being assessed and that it does not give a positive result because of other conditions.

The ROC curve of our method is close to the left-hand border and then the top border of the ROC space. On the other hand, ROC curves of AM method and SVMs method are close to the 45-degree diagonal of the ROC space. The ROC curve demonstrates that our method has a considerably better performance than those of AM and SVMs method.

In the test with negatives examples chosen in separate sub-cellular compartments, we carried out 10 trials of 10-folds cross validation, and then calculated the average sensitivity (SS) and specificity (SP) of these 10 trials. Our method outperformed with sensitivity 84% and specificity 90% in the comparison with AM method with SS 82% and SP 34%, and SVMs method with SS 47% and SP 75%.

Reproducing the same experiments to non domain-based approach using ILP[11] with the same training negatives (with different numbers of negatives) and positives (Ito *et al.*'s date set of at least 3 hit interactions), the results of 10 times 10-fold cross-validation demonstrated in Figure 2, show that our integrative domain-based method achieved higher sensitivity, and higher or equal specificity, than the non-domain based approach.

Furthermore, the unknown interacting protein pairs are in fact much larger than known ones, we also did the comparative experiments with imbalance training sets. According to,[31] the negative example set should be 4 times larger than the positive example set, we random selected 2,500 positives from DIP core data set and random 10,000 negatives. Sensitivity and specificity of our method are 78% and 95% (in this case, SS and SP of AM are 75% and 30% respectively, and SS and SP of SVMs are 30% and 94%, respectively). As the result, even testing with imbalanced training data sets, our methods is effectively predict PPI.

3.2. *Predicting domain-domain interactions*

3.2.1. *Experiment design of domain-domain interactions prediction*

Similar to protein-protein interactions, domains of proteins interact together to perform specific biological function in cell. Inheriting the ILP framework for PPI prediction, we applied ILP framework to infer domain-domain interactions. Combining different databases, both databases of proteins and databases of domains, the results of experiments for DDI prediction are promising.

To assess the performance of our method for DDI prediction, sensitivity and specificity were evaluated through the 10-fold cross validation tests. We used about 3000 interaction in InterDom database as positive examples.[33] Positive examples are domain-domain interactions in InterDom database that have score thresholds over 100 and are not false positives. Because there are currently no efficient experimental and computational methods for detecting non-interacting domain pairs, the negative examples were randomly generated. A domain pair is considered to be a negative example, if the pair does not exist in the interaction set. Various numbers of negatives, i.e. 500, 1000, 2000, 3000 negatives, were chosen.

3.2.2. *Result of domain-domain interactions prediction*

In fact, the interaction of two domains depends on: (i) domain features of interacting partners themselves, and (ii) protein features of host proteins consisting of those domains. In case of domain-domain interaction prediction, we did not use domain-domain interaction data of iPfam database and domain fusion data in ILP background knowledge.

We modeled 20 predicates from 7 databases (see more in Supplementary materials[34]). Among these 20 predicates, there are 14 predicates as protein features extracted from 3 genomic/proteomic databases such as UniProt database, CYGD database, and GO database are protein features in forms of 14 predicates. In addition, there are 7 predicates for domain features corresponding to 4 domain databases, i.e. Pfam (http://www.sanger.ac.uk/Software/Pfam/), PRINT (http://www.bioinf.manchester.ac.uk/dbbrowser /PRINTS/), and PROSITE (http:// au.expasy.org/prosite/), and Interpro (http://www.ebi.ac.uk/interpro/). With more than 100,000 ground facts of 20 predicates extracted from 7 databases, we efficiently predict domain-domain interactions by ILP. The target predicate for DDI prediction is `interact_domain(+protein, +protein)`.

Results conducted from 10 times of 10-fold cross-validation showed that our method obtains high sensitivity and specificity. The performance of our method in terms of specificity and sensitivity is also statistically tested by confidence intervals. The confidence intervals give us an estimate of the amount of error involved in our data. To estimate 95% confidence interval for each calculated specificity and sensitivity, we used t distribution. The 95% confidence intervals of specificity and sensitivity are shown in Table 2.

Table 2. The sensitivity and specificity are obtained for each randomly chosen set of negative examples by 10 times 10-fold cross-validation.

# Negative example	Sensitivity	Specificity
500	0.83 ± 0.016	0.61 ± 0.075
1000	0.78 ± 0.042	0.68 ± 0.042
2000	0.69 ± 0.027	0.80 ± 0.018
3000	0.73 ± 0.028	0.84 ± 0.010

Besides calculating cross-validated sensitivity and specificity, cross-validated accuracy and precision are considered. All of our experiment results had high accuracy and precision. The average accuracy and precision of our method were 0.76 and 0.82, respectively.[34]

4. Discussion

The experimental results have shown that ILP approach potentially predicts PPI and DDI with high sensitivity and specificity. Furthermore, the inductive rules of ILP encouraged us to discover many interesting biological reciprocal relationships among protein-protein interactions and protein domains, and other genomic/proteomic features related to protein-protein interactions. Analysing our results in comparison with information in biological literatures, we found that ILP induced rules could be further applied to related studies in biology. The list both of PPI prediction rules and DDI prediction rules are available as Supplemental materials.[34] The following section is the analysis of some of these rules.

Studying the rules of PPI prediction related to domain-domain interaction information, we found many interesting rules. For example, the following rule shows that if two proteins, A and B, have domains belonging to domain databases like PROSITE or InterPro (having a PROSITE or InterPro annotation C) and have at least one mediating domain-domain interaction, they may interact

has_int (A,B) :- dr_prosite (B, C), dr_prosite (A, C), ddi (A, B, yes)

with 43 positives covered

has_int(A,B) :- dr_interpro(B,C), dr_interpro(A,C), ddi (A, B, yes)

with 90 positives covered.

A large number of positives, which indicates these rules, confirms why domain-domain interactions are considered as key factors to predict PPI.

Considering the group of proteins which may be required for the production of *pyridoxine* (vitamin B6) sno1_yeast, snz3_yeast snz1_yeast, and snz2_yeast, we found that each pair in this group has an interaction which satisfies the following rule

has_int(A,B) :- ig (A, B, C), C = 1, ddi (A, B, yes),
function_cat (B, cell rescue defense and virulence).

This rule means interaction of protein A and protein B may occur if the proteins satisfy three conditions. First is that they interact with the same protein. Second is that they have at least one DDI. Third is that one of them is categorized to function catalogue 'cell rescue defense and virulence'. We knows that PPI plays an important role in drug design, so such rules and their evidence are expected to help us to discover interesting relationships between PPI, DDI and protein function in pharmaceuticals.

Two most popular rules related to domain fusion information are

has_int(A,B) :- dr_go(B,C), part_of(C,D), domain_fusion(A, B,yes)
has_int(A,B) :- dr_go(B,C), dr_go(A,C), domain_fusion(A,B,yes)

The first one covers 199 positives and the second one covers 217 positives. Both of these rules consist of GO term and domain fusion information. According to the second rule, if two proteins have GO terms and their domains are fused in another protein, there may occur an interaction.

Our induced rules with large number of positives prove that if a pair of proteins, A and B, are located in the same subcellular compartment, protein A potentially interacts with protein B. There are 216 covered positives for 'nucleus compartment', 284 ones for 'cytoplasm compartment', and 15 ones for 'mitochondria compartment'. However, surprisingly among induced rules, we found a rule with 37 positives that showed the phenomenon of two proteins being in different subcellular locations but interacting

has_int(A,B) :- subcell_cat(B,nucleus), subcell_cat(A,cytoplasm),
function_cat(A,transcription).

This phenomenon could occur when there is a certain translocation or post-translation modification of proteins in different subcellular compartments.

Analysing DDI prediction rules, some interesting associations between DDI and other domain and protein features are discovered.

Related to *motif compound* feature in domain, we found that the more motifs a domain has, the more interactions the domain has with other domains. This means that domains which have many conserved motifs tend to interact with others. The interactions of these domains play an important role in forming stable domain-domain interactions in particular, and protein-protein interactions in general.[35] If two domains, A' and B', the domain A' has a PRINTS entry C, and C is with eight motifs and the rest domain B' belongs to proteins categorized in *protein synthesis* function category, they interact. This rule covers 23 positives

interact_domain (A',B') :- prints (A', C), motif_compound
(C, compound(8)), function_category (B', protein synthesis).

The combination of inductive rules of ILP will be very useful for not only understanding PPI and DDI, but also protein functions, and biological processes.

5. Conclusion

We have presented an integrative domain-based approach using ILP and multiple genome databases to predict protein-protein interactions. The experimental results demonstrated that our proposed method could produce comprehensible rules and perform well in comparison with other methods on protein-protein interaction prediction. In future work, we would like to further investigate the biological significance of novel protein-protein interactions obtained by our method, and apply the ILP approach to other important tasks, such as determining protein functions, and determining the sites, and interfaces of PPI using DDI data.

6. Acknowledgements

The authors would like to thank Dr. Tuan Nam Tran and Professor Kenji Satou for helpful discussions and comments.

References

1. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y, A comprehensive two-hybrid analysis to explore the yeast protein interactome. In *Proc. Natl. Acad. Sci. USA 98*, pp. 4569–4574, 2001.
2. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM, A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature*, **403**(6770):623–627, February 2000.
3. Bauer A, Kuster B, Affinity purification-mass spectrometry: Powerful tools for the characterization of protein complexes. *Eur. J. Biochem.*, **270**(4):570–578, 2003.
4. Smith GP, Filamentous fusion phage: Novel expression vectors that display cloned antigens on the virion surface. *Science*, **228**(4705):1315–1317, 1985.
5. Bock JR, Gough DA, Predicting protein – protein interactions from primary structure. *Bioinformatics*, **17**(5):455–460, 2001.
6. Matthews LR, Vaglio P, Reboul J et al., Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or 'interologs'. *Genome Res.*, **11**(12):2120–2126, 2001.
7. Pellegrini M, Marcotte EM, Thompson MJ et al., Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. In *Proc. Natl. Acad. Sci. USA*, **96**:4285–4288, 1999.
8. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Mark G, A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data. *Science*, **302**(5644):449–453, 2003.
9. Ben-Hur A, Noble WS, Kernel methods for predicting protein-protein interactions. *Bioinformatics*, **21**(suppl1):i38–46, 2005.
10. Zhang LV, Wong SL, King OD, Roth FP, Predicting co-complexed protein pairs using genomic and proteomic data integration. *BMC Bioinformatics*, **5**(38), 2004.
11. Tran TN, Satou K, Ho TB, Using inductive logic programming for predicting protein-protein interactions from multiple genomic data. In *PKDD*, pp. 321–330, 2005.
12. Rhodes DR, Tomlins SA, Varambally S, Mahavisno V, Barrette T, Kalyana-Sundaram S, Ghosh D, Pandey A, Chinnaiyan AM, Probabilistic model of the

human protein-protein interaction network. *Nat Biotech*, **23**(8):1087–0156, 2005. http://www.nature.com/nbt/journal/v23/n8/suppinfo/nbt1103_S1.html.

13. Sprinzak E, Margalit H, Correlated sequence-signatures as markers of protein-protein interaction. *Journal of Molecular Biology*, **311**(4):681–692, 2001.
14. Kim RM, Park J, Suh JK, Large scale statistical prediction of protein - protein interaction by potentially interacting domain (PID) pair. In *Genome Inform. Ser. Workshop Genome Inform*, pp. 48–50, 2002.
15. Han D, Kim HS, Seo J, Jang W, A domain combination based probabilistic framework for protein protein interaction prediction. In *Genome Inform. Ser. Workshop Genome Inform*, pp. 250–259, 2003.
16. Wojcik J, Schachter V, Protein-protein interaction map inference using interacting domain profile pairs. *Bioinformatics*, **17**(suppl1):S296–305, 2001.
17. Chen XW, Liu M, Prediction of protein-protein interactions using random decision forest framework. *Bioinformatics*, **21**(24):4394–4400, 2005.
18. Martin S, Roe D, Faulon JL, Predicting protein-protein interactions using signature products. *Bioinformatics*, **21**:218–226, 2005.
19. Pawson T, Raina M, Nash N, Interaction domains: from simple binding events to complex cellular behavior. *FEBS Letters*, **513**(1):2–10, 2002.
20. Nguyen TP, Ho TB, Combining domain fusions and domain-domain interactions to predict protein-protein interactions. In *Proc. 7th International Workshop on Data Mining in Bioinformatics (BIOKDD '07)*, pp.27–34, 2007.
21. Muggleton S, *Inductive Logic Programming.* Academic Press, 1992.
22. Dzeroski S, Lavrac N, editors. *Relational Data Mining.* Springer, 2001.
23. Page D, Craven M, Biological applications of multi-relational data mining. In *SIGKDD Explorations*, volume 5, pp. 69–79, 2003.
24. King RD, Muggleton S, Lewis RA, Sternberg MJE, Drug design by machine learning: The use of inductive logic programming to model the structure activity relationships of trimethoprim analogues binding to dihydrofolate reductase. *Proc. of the National Academy of Sciences of the USA*, **89**(23):11322–11326, 1992.
25. Muggleton S, King RD, Sternberg MJE, Protein secondary structure prediction using logic-based machine learning. *Protein Eng.*, **6**(5):549–, 1993.
26. Turcotte M, Muggleton SH, Sternberg MJE, Protein fold recognition. In *Proc. of the 8th International Workshop on Inductive Logic Programming (ILP-98)*, pp. 53–64, 1998.
27. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D, Detecting Protein Function and Protein-Protein Interactions from Genome Sequences *Science*, **285**(5428):751–753, 1999.
28. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA, Protein interaction maps for complete genomes based on gene fusion events, *Nature*, **402**:86-90, 1999.
29. Truong K, Ikura M, Domain fusion analysis by applying relational algebra to protein sequence and domain databases. *BMC Bioinformatics*, **4**(16):1–10, 2003.
30. Srinivasan A, A learning engine for proposing hypotheses. `http://web.comlab.ox.ac.uk/oucl/research/areas/machlearn/Aleph/aleph.html`
31. Ben-Hur A, Noble WS, Choosing negative examples for the prediction of protein-protein interactions. *BMC Bioinformatics*, **7**(suppl 1):–, 2006.
32. Joachims T, Making large-scale support vector machine learning practical. In B. Scholköpf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods: Support Vector Machines.* MIT Press, Cambridge, MA, 1998.
33. Ng SK, Zhang Z, Tan SH, Lin K, InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes. *Nucleic Acids*

Res, **31**(1):251–254, 2003.

34. Nguyen TP, Ho TB, Supplementary materials, 2007. `http://www.jaist.ac.jp/~s0560205/PPIandDDI/`.
35. Moon HS, Bhak J, Lee KH, Lee D, Architecture of basic building blocks in protein and domain structural interaction networks. *Bioinformatics*, **21**(8):1479–1486, 2005.