Discovering Signal Transduction Networks Using Signaling Domain-Domain Interactions

Thanh Phuong Nguyen phuong@jaist.ac.jp
Tu Bao Ho bao@jaist.ac.jp

Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan

Abstract

The objective of this paper is twofold. One objective is to present a method of predicting signaling domain-domain interactions (signaling DDI) using inductive logic programming (ILP), and the other is to present a method of discovering signal transduction networks (STN) using signaling DDI.

The research on computational methods for discovering signal transduction networks (STN) has received much attention because of the importance of STN to transmit inter- and intra-cellular signals. Unlike previous STN works functioning at the protein/gene levels, our STN method functions at the protein domain level, on signal domain interactions, which allows discovering more reliable and stable STN. We can mostly reconstruct the STN of yeast MAPK pathways from the inferred signaling domain interactions, with coverage of 85%. For the problem of prediction of signaling DDI, we have successfully constructed a database of more than twenty four thousand ground facts from five popular genomic and proteomic databases. We also showed the advantage of ILP in signaling DDI prediction from the constructed database, with high sensitivity (88%) and accuracy (83%). Studying yeast MAPK STN, we found some new signaling domain interactions that do not exist in the well-known InterDom database. Supplementary materials are now available from `http://www.jaist.ac.jp/s0560205/STP_DDI/`.

Keywords: signal transduction network, signaling domain, signaling domain-domain interaction, protein-protein interaction, inductive logic programming

1 Introduction

Signal **T**ransduction **N**etworks (STN) are the primary means by which eukaryotic cells respond to external signals from their environment and coordinate complex cellular changes [2]. Because of the biologically significant roles of STN in the cell, both biologists and bioinformaticians have taken much interest in finding out molecular components and/or the relations among these molecular components in STN. These works were traditionally done by biological experimental methods such as gene knockouts and epistasis analysis. On the one hand, *in vitro* biological experiments usually require much effort and time. One the other hand, although the experimental methods are effective in the discovery of molecular components, the study of the relations among these molecular components is still challenging. With the accumulation of genome sequence information, large-scale genomic and proteomic techniques have offered insights into the components of signal transduction networks, as well as the molecular relations in signaling transduction networks [6]. Therefore, there is a great need to develop computational methods to direct biological discovery, enabling biologists to discover the mechanisms underlying complex signal transduction networks.

Consequently, discovering the relations among molecular components of signaling networks using large-scale genomic and proteomic information is an area of much ongoing research. A statistical model, based on representing proteins as collections of domains or motifs, which predicts unknown

molecular interactions within these biological networks was proposed by Gomez *et al.* [5]. Using Markov chain Monte Carlo method, they then modeled the signal transduction networks in terms of domains in upstream and downstream protein interactions. Steffen *et al.* [14] developed a computational approach for generating static models of signal transduction networks which utilizes protein-protein interaction maps generated from large-scale two-hybrid screens and expression profiles from DNA microarrays. Liu *et al.* [6] applied a score function that integrated protein-protein interaction data and microarray gene expression data to predict the order of signaling pathway components. Concerning protein modification time-course data, Allen *et al.* [2] applied a method of computational algebra to the modeling of signaling networks. These previous works sometimes generate lots of both false positives and false negatives, and/or are time-consuming. They also have two drawbacks related to the reliability and stability of STN. First, they consider molecular components of STN at protein or gene levels only, however there are other smaller basic units which actually transport signals in the cell. Second, they do not deal with phenomena underlying the interactions/relations among molecular components of STN.

The reliability and stability of STN depend much on signaling features of molecular components and the stability of the interactions or relations among these molecular components in the STN [1]. Within a protein, a domain is a fundamental structural unit that is self-stabilizing, and often folds independently of the rest of the protein chain. Domain-domain interactions are crucial in forming stable protein-protein interactions (PPI), and take part in many cellular processes and biochemical events. Signaling domains are often named and singled out because they figure prominently in the signaling features of STN. Signaling domains are primary units to transmit cellular signals with catalytic, adaptor, effector, and/or stimulator functions [12]. These signaling domains interact physically with one another to form stable channels in terms of protein-protein interactions or protein-protein relations, to send and receive intra-, extra-signals in STN. For example, *TIR* domain interactions between receptors and adaptors play a key role in activating conserved cellular signal transduction pathways in response to *bacterial LPS*, *microbial* and *viral pathogens*, *cytokines* and *growth factors*. Our key idea is to discover more reliable and stable STN by using the signaling DDI.

Understanding signaling domain-domain interactions (signaling DDI), we could grasp STN in depth in terms of both basic signaling units and the mechanism of signal transduction among proteins in STN. Recently, there are several works that have attempted to discover domain-domain interactions (DDI). An integrative approach is proposed by Ng *et al.* [10] to infer putative domain-domain interactions from three data sources, including experimentally-derived protein interactions, protein complexes and Rosetta stone sequences. The maximum likelihood estimation (MLE) is applied to infer the likelihood of domain-domain interactions by analyzing the observed protein interaction data [4]. Chen *et al.* [3] used domain-based random forest framework to predict DDI. Riley *et al.* [11] proposed a domain pair exclusion analysis (DPEA) for inferring DDI from databases of protein interactions. These works mostly exploited only one protein database but none of the individual protein databases can provide all information needed to perform better DDI prediction. Besides, domain-domain interactions depend on features of domains – not only features of proteins [1, 7]. Using only protein data (protein-protein interaction data in particular) is also one limitation of the previous works. These works predicted DDI in general, and did not consider properties of interacting partners, like signaling features of interacting domains.

In this paper, we propose a computational approach to discover more reliable and stable STN using signaling domain interactions. Our work solves two problems: (1) predicting authentic signaling DDI from diverse genomic and proteomic databases, (2) discovering STN using signaling DDI. For signaling DDI prediction, we first examine five most informative genome databases, and extract more than twenty four thousand possible and necessary ground facts on signaling protein domains. We then employ inductive logic programming (ILP) to infer efficiently signaling DDI. Sensitivity (88%) and accuracy (83%) obtained from 10-fold cross validation show that our method is useful for predicting signaling domain interactions. Studying yeast MAPK pathways, we predicted some new signaling DDI

that do not exist in the well-known InterDom database. Assuming all proteins in STN are known, we preliminarily build up signal transduction networks between these proteins based on their signaling domain interaction networks. We can mostly reconstruct the STN of yeast MAPK pathways from the inferred signaling domain interactions with coverage of 85%. Our approach could easily and flexibly be applied with other organisms, and various genomic and proteomic databases as well.

2 Method

In this section, we describe our proposed method to discover STN using signaling DDI. Two main tasks of the method are: (i) Applying ILP to predict signaling domain interactions from multiple genomic and proteomic databases, and (ii) Discovering STN using signaling DDI.

2.1 Predicting Signaling Domain Interactions Using ILP

Inductive Logic Programming (ILP) is an area of AI built on a foundation laid by research in machine learning and computational logic. ILP deals with the induction of hypothesized predicate definitions from examples and background knowledge. ILP is differentiated from most other forms of Machine Learning both by its use of an expressive representation language, and by its ability to make use of logically-encoded background knowledge. This has allowed successful applications of ILP in areas such as molecular biology and natural language, which both have rich sources of background knowledge, and both benefit from the use of expressive concept representation languages [8]. ILP is particularly suitable for bioinformatics tasks because of its ability to take into account background knowledge and work directly with structured data. ILP has been applied to many tasks in bioinformatics, such as protein secondary structure prediction [9] protein fold recognition [16], and protein-protein interaction prediction [15].

Algorithm 1 Predicting signaling domain-domain interactions from multiple genomic and proteomic databases.

Input:

The set D of domain-domain interactions extracted from InterDom database.
Signaling domains set S extracted from [19] and [20] .
Number of negative examples $\neg d_{ij}$ N.
Multiple genomic and proteomic databases Interpro, PRINTS, Uniprot, MIPS, GO database denoted by $(S^{InterPro}, S^{PRINTS} S^{Uniprot}, S^{MIPS}, S^{GO})$.

Output: Set of rules R for signaling domain-domain interaction prediction.

1: $R := \emptyset$.
2: $I := \emptyset$. {I is the set of positive examples $d_{ij}s$}
3: **for each** $d_i \in S$
4: **for each** $d_j \in S$
5: **if** $(d_i, d_j) \in D$ **then**
6: $I = I \cup \{d_{ij}\}$
7: Generate negative examples $\neg d_{ij}$s by selecting randomly N domain pairs $(d_i, d_j) \in S$ where $(d_i, d_j) \notin I$.
8: **for each** domain $d_i \in I$
9: Extract data from genomic/proteomic databases M $(\forall M \in (S^{InterPro}, S^{PRINTS}, S^{Uniprot}, S^{MIPS}, S^{GO}))$.
10: Integrate and formalise all extracted data in Aleph background knowledge language.
11: Select a positive example d_{ij} at random.
12: Saturate it to find the most specific clause that entails this example.
13: Do top-down search for selecting the best clause c.
14: $R := R \cup \{c\}$.
15: Remove covered positive examples.
16: **if** there remain positive examples **then goto** Step 11.
17: **return** R.

Algorithm 1 describes the proposed ILP framework for predicting signaling DDI from multiple genomic and proteomic databases. In this paper, we applied Aleph system (A Learning Engine for Proposing Hypotheses) [22] to learn background knowledge and induce rules. Among many ILP systems, Aleph is advanced and flexible in that it allows customisation of search, cost functions, output-display, etc. Aleph uses a top-down ILP covering algorithm as default, taking as input background information in the form of predicates, a list of modes declaring how these predicates can be chained together, and a designation of one predicate as the head predicate to be learned. Aleph is able to use a variety of search methods to find good clauses, such as the standard methods of breadth-first search, depth-first search, iterative beam search, as well as heuristic methods requiring an evaluation function.

In Algorithm 1, Steps 2 to 6 are used to generate positive examples, and Step 7 randomises negative examples (see more in Section 3.1). In Steps 8 to 10, there are two tasks: (1) extracting data from multiple genomic and proteomic databases, and (2) integrating extracted data, then formalising data in ground fact terms[1], as restricted in Aleph system. The procedure from Step 11 to Step 16 is the Aleph procedure to learn three input files (background file, positive examples file, and negative examples file). In this phase, we use the default evaluation function *coverage* (the number of positive and negative examples covered by the clause). The output is the set of predictive rules R. The discussion of some output rules appears in Section 4.2.

2.2 Discovering Signal Transduction Networks Using Signaling DDI

STN are usually represented by sets of molecular components like gene products, mostly proteins including RNAs. Between these proteins, there are protein interactions/relations/bindings/associations to transduce cellular signals.

Algorithm 2 Discovering signal transduction networks using signaling domain-domain interactions.

Input: Set of all proteins p_ks P in a signal transduction network Ω.
Output: A signal transduction network Ω in terms of a set of signaling domain interactions $S^{interact}$.

1: $K := \emptyset$. {Set K is the set of domains d_is belonging to proteins p_ks}
2: $T := \emptyset$. {Set T is the set of domain pairs (d_i, d_j)s where d_i, d_j belong to proteins p_ks}
3: **for all** proteins p_ks $\in P$
4: Extract all domains d_is belonging to the proteins p_ks.
5: **if** $d_i^k \notin K$ **then** $K = K \cup \{d_i^k\}$.
6: **for all** domains d_i^ks of proteins p_ks and domains d_j^ls of proteins p_ls
7: Couple the domains d_i^ks with domains d_j^ls $\forall p_k, p_l \in P$ and $p_k \neq p_l$.
8: $T = T \cup \{(d_i^k, d_j^l)\}$.
9: **call** Algorithm 1 with testing set T to predict signaling domain interactions d_{ij}^{kl}s; $S^{interact} = S^{interact} \cup \{d_{ij}^{kl}\}$.
10: **for all** protein pairs (p_k,p_l)
11: **if** (p_k,p_l) having at least one signaling domain interaction $d_{ij}^{kl} \in S^{interact}$ **then**
12: Connect protein p_k with protein p_l to form an edge in the STN Ω.
13: Estimate W_Ω^{STN} for the STN Ω.
14: **return** STN Ω and W_Ω^{STN}.

We considered an STN as a network of proteins. In the network, each protein is a node, and interactions, and relations (or bindings/associations) are the edges. Underlying these edges, there are (signaling) domain-domain interactions which are key channels to send and/or receive cellular signals among proteins. We simply assume that if one protein tends to "contact" another to transduce signals, at least one (signaling) domain interaction is required. Algorithm 2 demonstrates our proposed method to discover STN from (signaling) DDI, given that all proteins in STN are known.

Based on protein-protein interaction information, the reliability score of STN is proposed as follows:

[1] The term 'ground fact' is used here as in inductive logic programming.

$$W_{\Omega}^{STN} = L_{\Omega} * S_{\Omega} \prod_{i,j=1} \left(w_{ij} + \frac{\theta_{ij}}{\sum\limits_{i,j=1} \theta_{ij}} \right), \quad (1)$$

where

w_{ij}: the weight of DDI d_{ij} in terms of the frequency in a STN Ω.

L_{Ω}: the ratio of the number of predicted DDI over the number n DDI expected in the signal transduction network Ω.

S_{Ω}: the ratio of the number of predicted signaling DDI over the number n DDI expected in signal transduction network Ω.

θ_{ij}: number of protein-protein interactions containing the DDI d_{ij}.

Equation (1) evaluates the reliability of inferring the STN from the predicted signaling DDI. The evaluation is presented in Section 4.

3 Materials

3.1 Training Datasets

This paper concentrates on predicting DDI for *Saccharomyces cerevisiae* – a budding yeast, as the *Saccharomyces cerevisiae* data is available. The set of signaling domains is extracted from *SMART* database [20], and from the scientific literature collected by Pawson and his colleagues [19], denoted by S. For example, domain *sh2*, and domain *sh3* are key elements for transmitting signals in cells. After excluding the overlapping signaling domains, set S consists of 100 signaling domains.

InterDom database is the well-known DDI database consisting of more than 37,000 domain-domain interactions of multiple organisms. However, for signaling domain interactions, there is not any available database yet. Then, the set of positive examples I is obtained as the set of signaling domain pairs (d_i, d_j) in [19, 20] having an interaction in InterDom database. Each domain d_i is coupled with another d_j, and a signaling domain pair (d_i, d_j) is called a positive example if this pair is found in the InterDom database (see Steps 2 to 4 in Algorithm 1). Through this procedure, the final positive set I consists of 472 signal domain interactions. Also, the database of non domain-domain interaction does not yet exist, so the negative examples $\neg d_{ij}$s are generated at random (see Step 6 in Algorithm 1). In the experiment in this paper, we chose randomly 100 negatives $\neg d_{ij}$.

3.2 Genomic and Proteomic Datasets for Generating ILP Ground Facts

Unlike previous work mentioned in Section 1, we chose and extracted data from five genomic and proteomic databases to generate background knowledge[2] with an abundant number of ground facts, and used this data to predict signaling DDI. The five genomic and proteomic databases used are *(i) Protein fingerprints database* (PRINTS) [21], *(ii) Protein families and domains database* (InterPro) [23], *(iii) Universal Protein Resource* (Uniprot) [25], *(iv) The Mammalian Protein-Protein Interaction Database* (MIPS) [18], and *(v) Gene Ontology* (GO) [24].

Aleph uses *mode declarations* to build the bottom clauses, and there are three types of variables: (1) the input variable (+), (2) the output variable (−), and (3) the constant term (#). In this paper, target predicate is `domain_interaction(domain, domain)`. The instances of this relation represent the interaction between two signaling domains. For background knowledge, all domain/protein data are shortly denoted in form of different predicates. Table 1 shows the list of predicates used as background knowledge for each data source. From each data source, the predicates present some features thought to be useful for signaling DDI prediction. For example, pred-

[2] The term 'background knowledge' is used here as in inductive logic programming.

Table 1: Predicates used as background knowledge generated from genomic and proteomic data sources.

Data source	Background knowledge predicates	
InterPro	`interpro2go(+InterPro_Domain,-GO_Term)` Mapping of InterPro entries to GO	`interpro(+Domain,-InterPro_Domain)` A domain has a InterPro annotation number
PRINTS	`motif_compound(+Domain,#motif_compound)` A domain belongs to proteins having a number of motifs	`prints(+Domain,-PRINTS_Domain)` A domain has a PRINTS annotation number
Uniprot	`haskw(+Domain,#Keyword)` A domain has keywords of its proteins	`hasft(+Domain,#Feature)` A domain has features of its proteins
	`ec(+Domain,#EC)` A domain has coded enzymes of its proteins	`pir(+Domain,-PIR_Domain)` A domain has a PIR annotation number
	`biocyc(+Domain,#BioCycle)` A domain has Biocycle annotations of its proteins.	
GO	`is_a(+GO_Term,-GO_Term)` `is_a` relation between two GO terms	`part_of(+GO_Term,-GO_Term)` `part_of` relation between two GO terms
	`go(+Domain,-GO_Term)` A domain has GO terms of its proteins	
MIPS	`subcellular_location(+Domain,#Subcellular_Structure)` A domain has subcellular structures in which its proteins are found.	
	`function_category(+Domain,#Function_Category)` A domain has the proteins categorized to certain function categories	
	`domain_category(+Domain,#Protein_Category)` A domain has proteins categorized to certain protein categories	
	`phenotype_category(+Domain,#Phenotype_Category)` A domain has proteins categorized to certain phenotype categories	
	`complex_category(+Domain,#Complex_Category)` A domain has proteins categorized to certain complex categories	
Others	`num_int(+Domain,#num_int)` A domain has a number of interactions	`ig(+Domain,+ Domain, #ig)` A domain has interaction generality

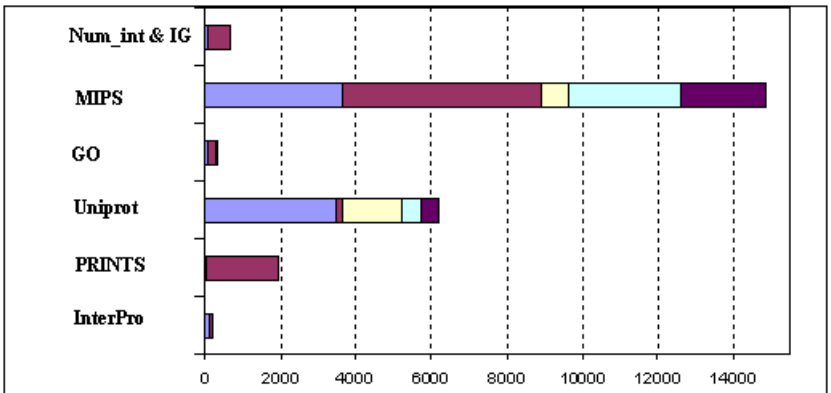
icate `motif_compound(+Domain,#motif_compound)` is predictive for signaling DDI prediction and gives the information about the stability of signaling DDI. Concerning on functions of host proteins of domains, predicate `function_category(+Domain,#Function_Category)` is generated. Proteins in a signaling transduction network should join together in a protein complex, so predicate `complex_category(+Domain,#Complex_Category)` presents the relation between one domain and some complex categories of its proteins. (For example, *complex_category(pf00400, transcription_complexes)* where *pf00400* is Pfam accession number and *transcription_complexes* is complex category name.) Other predicates are also generated from MIPS database, Uniprot database, GO database, and InterPro database.

Each predicate in Table 1 has many ground facts extracted from five genomic and proteomic data sources. Figure 1 illustrates the number of ground facts obtained from each data source. The data table in this figure shows the number of ground facts extracted from databases for individual predicates demonstrated in Table 1. For example, with predicate

$$\texttt{subcellular\_location}(+\texttt{Domain}, \#\texttt{Subcellular\_Structure})$$

extracted from MIPS database, there are 3,616 ground facts extracted. With the nineteen background predicates, we obtained 24,123 ground facts in total associated with signaling DDI prediction.

Among five data sources, MIPS database (the latest update version in 2006) is outstanding, with 14,865 ground facts (more than 50%). A total of 6,207 ground facts are extracted from Uniprot database. The number of ground facts from these two databases is huge, because of the availability of these databases. Also, the relationships between domains and their proteins are many-many relationships. One domain can belong to many proteins, for example, domain *SH3* is found in 24 proteins of *Saccharomyces cerevisiae*, some of which are *Nuclear fusion protein FUS1*, *SSU81 protein (SHO1 osmosensor)* or *Cytokinesis 2 protein*, etc. And one protein can contain many domains, for example, protein *sln1_yeast* has three domains *HATPase_c, HisKA* and *REC*. Then, one domain has lots of ground facts for one predicate.



	Interpro	PRINTS	Uniprot	GO	MIPS	Num_intIG
Predicate 5			467		2274	
Predicate 4			533		2968	
Predicate 3			1543	29	709	
Predicate 2	48	1900	187	217	5298	571
Predicate 1	119	43	3477	65	3616	68

Figure 1: Number of ground facts extracted from each data source.

On the other hand, when extracting ground facts from other databases like PRINTS database, Interpro database, etc., the relationships between domains and their entries in these databases are one-one relationships or one-many relationships. Therefore, the number of ground facts extracted from those databases is less than from MIPS database and Uniprot database. The combination of all ground facts generated from five genomic and proteomic databases constructed considerable background knowledge associated with signaling DDI.

4 Experimental Results and Discussion

4.1 Prediction of Signaling Domain-Domain Interactions

To validate our proposed method of predicting signaling domain-domain interactions, we conducted a 10-fold cross-validation test. Our experimental results obtained higher sensitivity, specificity, accuracy and precision compared with AM method [13]. The sensitivity of a test ($\frac{TP}{TP+FN}$) is described as the proportion of true positives (TP) it detects of all the positives (true positives (TP) + false negatives(FN)), measuring how accurately it identifies positives. On the other hand, the specificity ($\frac{TN}{TN+FP}$) of a test is the proportion of true negatives (TN) it detects of all the negatives (true negatives (TN) + false positives(FP)), and thus is a measure of how accurately it identifies negatives. Besides sensitivity and specificity, accuracy ($\frac{TP+NP}{TP+FP+TN+FN}$) and precision ($\frac{TP}{TP+FP}$) were evaluated. The results of the 10-fold cross-validation test are shown in Figure 2.

The lower bound on the number of positive examples to be covered by an acceptable clause is 3, and no negative examples (noises) allowed to be covered by an acceptable clause. We obtained high sensitivity (88%), accuracy (83%), and precision (90%) compared with AM method (with sensitivity (50%), accuracy (50%) and precision (83%)). Because of the lack of one standard database of non-signaling domain interactions, the specificity of ILP method is not very high at 64% (52% with AM method).

The experimental results have shown that ILP approach potentially predicts DDI with high sensitivity and accuracy. Actually, we expect that the results will be much better when the datasets are more complete. Furthermore, the inductive rules of ILP encouraged us to discover lots of compre-

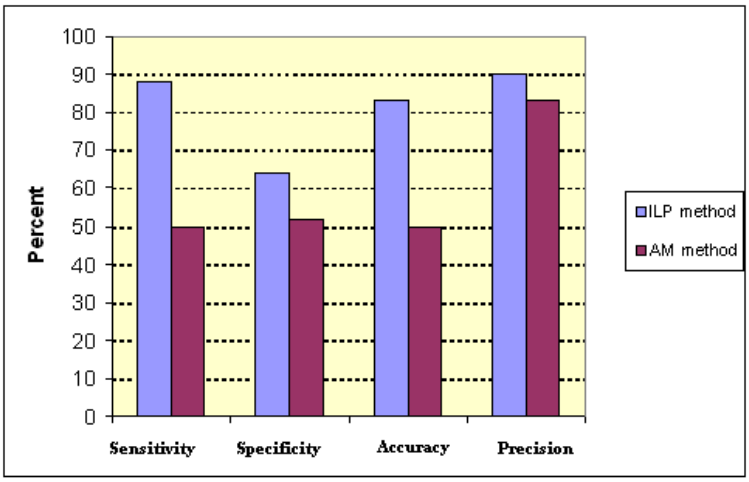


Figure 2: Performance of ILP method ($minpos = 3$ and $noise = 0$) compared with AM methods.

hensive relations between signaling DDI and protein features and, between signaling DDI and domain features. The following are some induced rules by Aleph.

Rule 1 [Pos cover = 31 Neg cover = 0]
$domain\_interaction(A, B) :- \quad dr\_go(A, C), ec(B, ec2\_3\_1),$
$function\_category(A, cellular\_transport\_facilitation\_and\_transport\_routes).$
Rule 2 [Pos cover = 27 Neg cover = 0]
$domain\_interaction(A, B) :- \quad dr\_prosite(A, C), ec(A, ec3\_1\_3), function\_category(B, cellular\_communication).$
Rule 3 [Pos cover = 20 Neg cover = 0]
$domain\_interaction(A, B) :- \quad prints(B, C), motif\_compound(C, compound(3)),$
$protein\_category(A, gtp - binding\_proteins).$
Rule 4 [Pos cover = 15 Neg cover = 0]
$domain\_interaction(A, B) :- \quad num\_int(A, C), C = 9, complex\_category(B, intracellular\_transport\_complexes).$

Rule 1 shows that if we have two domains, one of them belonging to proteins having the GO term and categorized in the $cellular\_transport\_facilitation\_and\_transport\_routes$ function category, and the other one belonging to proteins having coded enzyme $ec2\_3\_1$, then the two domains interact. Related to the $motif\_compound$ feature, we found there are many induced rules combining the motif features and other protein features. This means that the inferred interactions of these domains play an important role in forming stable protein-protein interactions in particular and stable STN in general [7]. Rule 3 is an example of such a rules. In Rule 3, if one domain has an annotation in PRINTS database, and the PRINTS annotation contains a compound of three motifs, that domain should interact with the other domain belonging to the proteins categorized in the $gtp - binding\_proteins$ category.

We found many comprehensive relations between signaling DDI and different domain and protein features from 74 induced rules. We expect that the combination of these rules will be useful for understanding signaling DDI in particular and protein-protein interactions, and signal transduction networks in general.

4.2 Discovering STN Using Signaling DDI

From predicted (signaling) domain interaction networks, we raise the question of how completely they cover the STN, and how to reconstruct STN using signaling DDI. Our motivation was to propose a computational approach to discover more reliable and stable STN using signaling DDI. When studying yeast MAPK pathways, the results of our work are considerable.

All extracted domains of proteins in MAPK pathways are inputs (testing examples) in our proposed predictor using ILP method (see Sections 2 and 3). With 32 proteins appearing in MAPK pathways, we extracted 29 different protein domains, and some of them are shared among proteins. Some

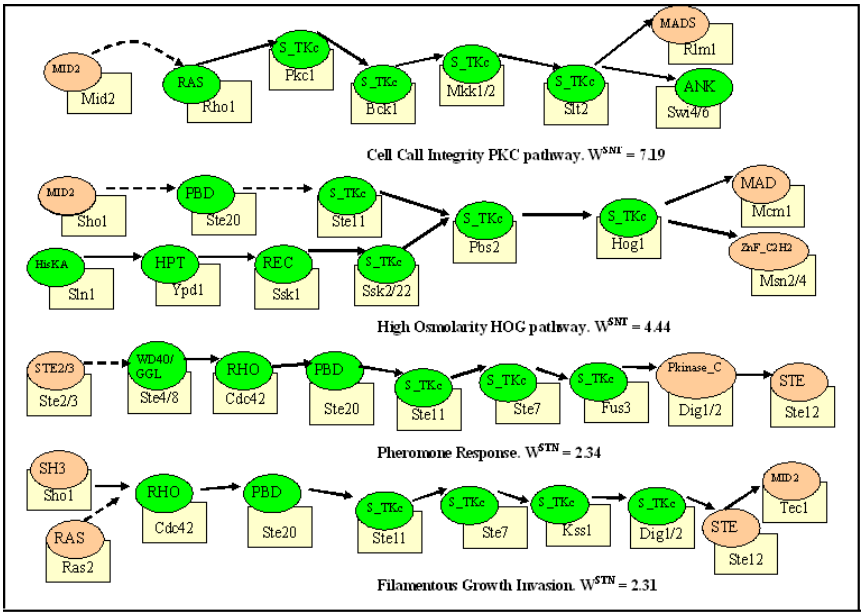


Figure 3: MAPK signal transduction pathways in yeast covered by signaling DDI networks. The rectangles denote proteins, the ellipses illustrate their domains and the signaling domains are depicted in dark. The signaling DDI are the lines with arrows, the missing interactions are dashed lines with arrows.

domains are determined to be signaling domains, such as domain *pf00069* belonging to many proteins, for example, *ste11_yeast*, *fus3_yeast* or *pbs_2*, etc., and some of them are not signaling domains, such as *TEA* or *MID2*. Figure 3 shows yeast MAPK (mitogen-activated protein kinase) covered by signaling domain interactions. MAPK pathways involve pheromone response, filamentous growth, and maintenance of cell wall integrity pathways. Table 2 shows the results of predicted signaling DDI when reconstructing STN for the yeast MAPK pathways. Moreover, among predicted signaling DDI for yeast MAPK pathways, there are some DDI which are newly discovered, when compared with the InterDom database. For example, our predicted DDI *(pf00071,pf00768)*, *(pf00768,pf00069)*, *(pf00433,pf02200)* do not exist in the InterDom database.

Evaluating signaling domain interactions predicted from the testing set of MAPK domains, 88% of protein relations in the Cell Wall Integrity PKC pathway, the Pheromone Response pathway, and the Filamentous Growth pathway are covered, and the Invasion High Osmolarity HOG pathway has coverage of 80%. Outstandingly, lots of domain interactions are found in which their corresponding proteins interacted in DIP (Database of Interacting Proteins) [17] and/or in CYGD (Comprehensive Yeast Genome Database) [18], for example, seven signaling domain interactions in the Cell Wall Integrity PKC pathway belong to 39 protein-protein interactions in CYGD database, and also belong to 47 protein-protein interactions in DIP. For estimating the reliability of STN, the reliability score W^{STN} (see Equation 1) was calculated for yeast MAPK pathways. The reliability score of the Cell Wall Integrity PKC pathway is the highest with $W^{STN} = 7.19$.

Table 2: Results of predicted signaling DDI in the yeast MAPK pathways.

The yeast MARK pathways	Percentage of signaling DDI predicted	#CYGD PPI covered	#DIP PPI covered
Cell Wall Integrity PKC pathway	88%	39	47
Pheromone Response	88%	41	42
Filamentous Growth	88%	40	38
Invasion High Osmolarity HOG	80%	40	53

5 Conclusion and Future Work

We have presented an approach using ILP and multiple genome and proteomic databases to predict signaling domain-domain interactions. The experimental results demonstrated that our proposed method could produce comprehensible rules, and at the same time, performed well compared with AM method on prediction of signaling domain-domain interaction. With a reliable set of predicted signaling domain interactions, the signal transduction networks were reliably discovered. A reliability score was proposed to estimate the reliability of STN inferred by signaling domain interaction networks. In future work, we would like to investigate further the biological significance of novel signaling domain-domain interactions obtained by our method. Selecting the features which are suitable to signaling domain interactions requires some further work. Also by applying graph techniques in data mining, it will be possible to model complete signal transduction pathways from signaling DDI networks.

References

[1] Albert, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J.D., *Molecular Biology of the Cell, 3rd Edition*, Garland Science, 1994.

[2] Allen, E.E., Fetrow, J.S., Daniel, L.W., Thomas, S.J., and John, D.J., Algebraic dependency models of protein signal transduction networks from time-series data, *J. Theor. Biol.*, 238(2):317–330, 2006.

[3] Chen, X.W., and Liu, M., Prediction of protein-protein interactions using random decision forest framework, *Bioinformatics*, 21(24):4394–4400, 2005.

[4] Deng, M., Mehta, S., Sun, F., and Chen, T., Inferring domain-domain interactions from protein-protein interactions, *Genome Res.*, 12(10):1540–1548, 2002.

[5] Gomez, S.M., Lo, S., and Rzhetsky, A., Probabilistic prediction of unknown metabolic and signal-transduction networks, *Genetics*, 159(3):1291–1298, 2001.

[6] Liu, Y. and Zhao, H., A computational approach for ordering signal transduction pathway components from genomics and proteomics data, *BMC Bioinformatics*, 5:158, 2004.

[7] Moon, H.S., Bhak, J., Lee, K.H., and Lee, D., Architecture of basic building blocks in protein and domain structural interaction networks, *Bioinformatics*, 21(8):1479–1486, 2005.

[8] Muggleton, S., Inductive logic programming: Issues, results and the challenge of learning language in logic, *Artificial Intelligence*, 114:283–296, 1999.

[9] Muggleton, S., King, R.D., and Sternberg, M.J.E., Protein secondary structure prediction using logic-based machine learning, *Protein Eng.*, 5(7):647–657, 1992.

[10] Ng, S.K., Zhang, Z., and Tan, S.H., Integrative approach for computationally inferring protein domain interactions, *Bioinformatics*, 19(8):923–929, 2003.

[11] Riley, R., Lee, C., Sabatti, C., and Eisenberg, D., Inferring protein domain interactions from databases of interacting proteins, *Genome Biol.*, 6(10):R89, 2005.

[12] Schultz, J., Milpetz, F., Bork, P., and Ponting, C.P., SMART, a simple modular architecture research tool: Identification of signaling domains, *Proc. Natl. Acad. Sci. USA*, 95(11):5857–5864, 1998.

[13] Sprinzak, E. and Margalit, H., Correlated sequence-signatures as markers of protein-protein interaction, *J. Mol. Biol.*, 311(4):681–692, 2001.

[14] Steffen, M., Petti, A., Aach, J., D'haeseleer, P., and Church, G., Automated modelling of signal transduction networks, *BMC Bioinformatics*, 3:34, 2002.

[15] Tran, T.N., Satou, K., and Ho, T.B., Using inductive logic programming for predicting protein-protein interactions from multiple genomic data. *9th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), Springer LNAI*, 3721:321–330, 2005.

[16] Turcotte, M., Muggleton, S.H., and Sternberg, M.J.E., Protein fold recognition, *Proc. of the 8th International Workshop on Inductive Logic Programming (ILP-98)*, 53–64, 1998.

[17] `http://dip.doe-mbi.ucla.edu`

[18] `http://mips.gsf.de/genre/proj/yeast/`

[19] `http://pawsonlab.mshri.on.ca`

[20] `http://smart.embl-heidelberg.de`

[21] `http://umber.sbs.man.ac.uk/dbbrowser/PRINTS/`

[22] `http://web.comlab.ox.ac.uk/oucl/research/areas/machlearn/Aleph/`

[23] `http://www.ebi.ac.uk/interpro/`

[24] `http://www.geneontology.org`

[25] `http://www.pir.uniprot.org`