Using Inductive Logic Programming for Predicting Protein-Protein Interactions – Some Preliminary Results

Tuan Nam Tran[†] and Tu Bao Ho[†]

Inductive Logic Programming (ILP) is differentiated from most supervised learning methods both by its use of an expressive representation language and its ability to make use of background knowledge. This has led to successful applications of ILP in molecular biology, such as predicting the mutagenicity of chemical compounds, predicting protein secondary structures, and discovering protein fold descriptions. In this paper, we attempt to apply ILP to the problem of predicting protein-protein interactions, which plays an essential role in bioinformatics since many major biological processes are controlled by protein interaction networks. We have used the Yeast Interacting Proteins Database provided by Ito, Tokyo University as training examples. Various kinds of background knowledge have been constructed by either extracting from protein databases or using computational approaches. Early results indicate that ILP is useful for obtaining comprehensible rules to differentiate those protein-protein interactions that are highly reliable. The predictive accuracy obtained using ten-fold cross-validation is nearly 80%, demonstrating a promising result of using ILP for predicting protein-protein interactions.

1. Introduction

The interaction between proteins is fundamental to a broad spectrum of biological functions, including regulation of metabolic pathways, immunologic recognition, DNA replication, progression through the cell cycle, and protein synthesis. The binding of one signaling protein to another can have a number of consequences. Firstly, such binding can serve to recruit to a signaling protein to a location where it is activated and/or where it is needed to carry out its function. Secondly, the binding of one protein to another can induce conformational changes that affect activity or accessibility of additional binding domains, permitting additional protein interactions. A cell in which suddenly the specific interactions between proteins disappear would become deaf and blind, paralytic and finally would disintegrate, because specific interactions are involved in almost any physiological process. Moreover, the study of protein-protein interactions plays an essential role in cancer treatment, providing important insight into the functions of many of the known oncogenes, tumor suppressors, and DNA repair proteins. Pharmacogenetic research has also expanded to include the study of drug transporters, drug receptors, and drug targets.

The full network of protein-protein interactions in model cellular systems should provide new insights into the structure and properties of these systems. An enormous amount of protein-protein interaction data have been obtained recently for yeast and other organisms using high-throughput experimental approaches such as yeast two-hybrid[14)], affinity purification and mass spectrometry[2)], phage display[29)]. However, a potential difficulty with these kinds of data is a prevalence of false positive (interactions that are seen in an experiment but never occur in the cell or are not physiologically relevant) and false negatives (interactions that are not detected but do occur in the cell). Although assessing the reliability of protein-protein interactions is an essential issue, there are still just a few studies that formulates mathematical measures. For example, Deng *et al.* have used a maximum likelihood method[10)] to estimate the reliability of different data sets. In this paper, we apply ILP to predicting protein-protein interactions with high reliability for the budding yeast *Saccharomyces cerevisiae*. We used the Yeast Interacting Proteins Database provided by Ito, Tokyo University as training examples since this data set contains a specific field called "IST hit" indicating how many times each in-

[†] School of Knowledge Science
Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Tatsunokuchi, Ishikawa 923-1292, Japan
email: {tt-nam,bao}@jaist.ac.jp

teraction was observed. Various kinds of background knowledge have been constructed by either extracting from protein databases or using computational approaches. Early results indicate that using ILP is a promising approach to differentiate protein-protein interactions that are highly reliable from a large amount of experimental data.

The remainder of this paper is organized as follows. Section 2 introduces some related work concerning protein-protein interactions. Section 3 focuses on ILP and the Progol system[19]. Section 4 describes our methodology and experimental result for predicting protein-protein interactions using Progol 5.0. Section 5 will summarizes our work and presents some future works.

2. Related work

This section describes some conventional methods applying computational approaches concerning protein-protein interactions. The computational methods concerning protein-protein interactions, in our view, can be classified into several groups: discovering protein-protein interactions, predicting and discovering knowledge on protein-protein interactions, and evaluating the reliability of protein-protein interactions. We will look at each group in detail.

2.1 Discovering PPI

Although there are many experimental methods for detecting protein-protein interactions, they suffer from many limitations such as holding high false positive and high false negative rates. Computational approaches in a rapid, automatic, and reasonably accurate manner would complement the experimental approaches. There exist many different computational approaches for screening entire genomes to discover protein-protein interactions from a variety of sources of information:

Based on structural homology Lu *et al.*[15] have used a threading-based algorithm in which we align the sequence of the protein of interest to a library of known folds to find the closest matching structure.

Based on interacting orthologs Matthews *et al.*[18] have investigated the extent to which a protein interaction map generated in one species can be used to predict interactions in another species under the interacting orthologs or "interologs" principle.

Based on gene neighborhood Dandekar *et al.*[4], based on the notion of conservation of gene neighborhood, have identified a number of genes which have been previously described to be physically interacting, showing that gene neighborhood is quite a powerful method for inferring protein-protein interactions in bacteria.

Based on gene fusion The study of Marcotte *et al.*[16] is based on the so-called Rosetta Stone[16] or gene fusion method, knowing that many genes become fused through the course of evolution due to selective pressure.

Based on phylogenetic profiles Pellegrini *et al.*[25] have constructed the phylogenetic profiles of proteins across a selected set of different genomes using the binary vector representation. In their study, the clusters of proteins formed by similar phylogenetic profiles tend to share the same functions comparing with random groups of proteins.

Based on phylogenetic tree similarity In a study by Goh *et al.*[12], they found a high correlation coefficient between the corresponding distance matrices of the two interacting protein domains, indicating that in a control set of known interactions, interacting protein pairs tend to have high correlation values in their distance matrices.

Based on (correlated) mRNA expression There is a significant relationship between gene expression and protein interactions on the proteome scale[13]. In fact, the mean correlation coefficients of gene expression profiles between interacting proteins are higher than those between random protein pairs.

For further readings on experimental and computational methods for discovering protein-protein interactions, the tutorial by Ng and Tan[21] provides a systematic review on these issues.

2.2 Predicting and discovering knowledge on PPI

There have been many studies using machine learning and data mining techniques for predicting and discovering knowledge on protein-protein interactions. Bock and Gough[3] have successfully applied a Support Vector Machine learning system to predict directly protein-protein interactions from primary structure and associated data. Deng *et al.*[9] predicted yeast protein-protein interactions using inferred domain-domain interactions, showing the interacting domain pairs can be useful for computational prediction of protein-protein interactions. Oyama *et al.*[24] have applied Association Rule Mining to extracting the knowledge from protein-protein interaction data.

On the other hand, there are a growing number of papers aim to extract protein-protein interactions from biomedical literature[28],[17].

3. ILP and Progol

Inductive Logic Programming (ILP) is the area of AI which deals with the induction of hypothesized predicate definitions from examples and background knowledge. Logic programs are used as a single representation for examples, background knowledge and hypotheses. ILP is differentiated from most other forms of Machine Learning (ML) both by its use of an expressive representation language and 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 an expressive concept representation languages[20].

The ILP is normally provided with background knowledge B, positive examples E^+ and negative examples E^- and constructs an hypothesis h. B, E^+ and E^- and h are each logic programs. A logic program is a set of definite clauses each having the form

$$h \leftarrow b_1, \ldots, b_n$$

where h is an atom and b_1, …, b_n are atoms. Usually E^+ and E^- consist of ground clauses, those for E^+ being definite clauses with empty bodies and those for E^- being clauses with head 'false' and a single ground atom in the body.

The conditions for construction of h are as follows.

Necessity: $B \not\models E^+$
Sufficiency: $B \wedge h \models E^+$
Weak consistency: $B \wedge h \not\models \Box$
Strong consistency: $B \wedge h \wedge E^- \not\models \Box$

The Sufficiency condition captures the notion of generalizing examples relative to background knowledge. A theorem prover cannot be directly applied to derive h from B and E^+. However, by simple application of the Deduction Theorem the Sufficiency condition can be rewritten as follows.

$$Sufficiency^* : B \wedge \bar{E^+} \models \bar{h}$$

This simple alteration has a profound effect. The negation of the hypothesis can now be deductively derived from the negation of the examples together with the background knowledge. This is true no matter what form the examples take and what form the hypothesis takes. This approach of turning an inductive problem into one of deduction is called *inverse entailment*[19]. Progol[19] is an ILP system based on inverse entailment, finding the most specific hypothesis (MSH) among all hypotheses with no conflict that can explain given positive examples.

4. Using ILP for predicting protein-protein interactions

We applied Progol 5.0 to predicting protein-protein interactions with high reliability. We used the Ito data set[7] to provide the training examples since this data set contains an attribute called IST hit, standing for how many times the corresponding interaction was observed. Intuitively, the higher IST hit, the much more reliable that interaction is. Ito *et al.*[14] conducted comprehensive analysis using their system to examine two-hybrid interactions in all possible combinations between the 6000 proteins of the budding yeast *Saccharomyces cerevisiae.*

Suppose E is a set of whole records with 4549 interactions, where each records is represented by e_i with the IST hit n_i. We can denote

$$E = \cup_{k=1}^{\infty} E_k$$

where

$$E_k = \{e_i \in E | n_i = k\}$$

. Given a IST hit threshold k, we divide into the set of positive (E_k^+) and negative (E_k^-) examples as below.

$$E_k^+ = \{e_i \in E | n_i > k\}$$

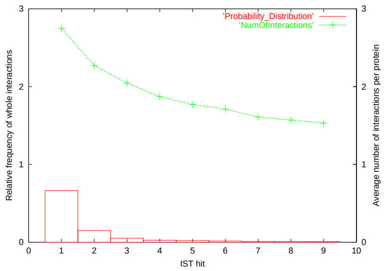

Fig. 1 Number of interactions for each IST hit among whole 4549 interactions

$$E_k^- = \{e_i \in E | n_i \le k\}$$

Figure 1 indicates the probability distribution and the average number of interactions per protein by varying the IST hit. Originally, Ito considers two kinds of data: *full data* with 4549 interactions, and *core data* consisting of 841 interactions (18.5%) with the IST hit is greater than two ($k = 2$). Note that the average number of interactions for a protein of the Ito data set is less than 3 and decreases when increasing the IST hit.

4.1 Preparing for training examples

The Ito data set consists of 4549 interactions, of which 841 interactions are positive (E_2^+), and 3708 interactions are negative examples (E_2^-). The original interactions are concerned with two ORFs (bait and prey ORF), however, in order to exploit the background knowledge in the SWISS-PROT database[6], we need to consider the related proteins, not ORFs. It should be noted that some ORFs occurred in the Ito data set may not occur in the SWISS-PROT database. Limited to the pairs of corresponding proteins occurring in the SWISS-PROT database, we obtained 592 positive and 2546 negative examples, respectively. The portion of positive examples before and after converting training examples is almost unchanged (18.5% and 18.9%, respectively). There are 2571 yeast proteins occurred in the training examples after converting.

4.2 Preparing for background knowledge

It should be noted that the key success to prediction of protein-protein interactions using ILP depends on how well we can provide the background knowledge. The background knowledge may be explicit, such as the location of a protein in the cell, or whether a protein contains a specific domain or not. In general, this information could be extracted from protein databases such as SWISS-PROT database[6]. This database can also be used as pointers to information related to entries and found in data collections other than SWISS-PROT. For this type of background knowledge, we have used the following predicates since they have a strong link to the functional property of proteins.

subcellular_location(proteinID,location) This predicate describes the location of the corresponding protein in the cell. For simplicity, we considered only three kinds of subcellular location including *nuclear*, *cytoplasmic*, and *mitochondrial*.

ec(proteinID,ec_category) This predicate describes the EC numbers concerning enzymes. Enzymes are classified based on their functions into four-level hierarchical categories, each of which is labeled by EC number. In the current work, we considered only three-level hierarchical categories, obtaining 87 EC categories.

dr_pir(proteinID,pir_category) This shows the link between the corresponding protein and entries in the PIR database[26]. We totally obtained 2534 PIR categories.

dr_interpro(proteinID,interpro_category) This shows the link between the corresponding protein and entries in the InterPro database[8]. We totally obtained 1404 InterPro categories.

dr_pfam(proteinID,pfam_category) This shows the link between the corresponding protein and entries in the Pfam database[11]. We totally obtained 1007 Pfam categories.

dr_prosite(proteinID,prosite_category) This shows the link between the corresponding protein and entries in the Prosite database[22]. We totally obtained 577 Prosite categories.

dr_go(proteinID,go_category) This shows the link between the corresponding protein and entries in the Gene Ontology[23]. We totally obtained 1333 GO categories.

On the other hand, the background knowledge can be implicitly provided using computational approaches. For example, we may use homology search to find proteins with similar functions. For this type of background knowledge, we have calculated the correlation coefficient between all protein pairs occurred in the positive and negative examples using the cell cycle gene expression data provided in[30]. This data contains 6,080 genes with 77 data points (2 cln3, 2 clb, 18 alpha, 24 cdc15, 17 cdc28, and 14 elut). We have used this predicate as the background knowledge due to the result obtained in[13] demonstrating the relation between the correlation coefficients of gene expression profiles and the interacting proteins.

1. $\mathit{high\_reliability\_ppi}(A, A) :- \mathit{dr\_interpro}(A, \mathit{ipr}003593).$
   A homotypic interaction occurs if the corresponding protein contains the InterPro domain "AAA_ATPase".
2. $\mathit{high\_reliability\_ppi}(A, A) :- \mathit{dr\_interpro}(A, \mathit{ipr}008631).$
   A homotypic interaction occurs if the corresponding protein contains the InterPro domain "Glycogen_synth".
3. $\mathit{high\_reliability\_ppi}(A, A) :- \mathit{ec}(A, \mathit{ec}2\_3\_1).$
   A homotypic interaction occurs if the corresponding protein contains the EC number 2.3.1 (transferring groups other than amino-acyl groups).
4. $\mathit{high\_reliability\_ppi}(A, A) :- \mathit{ec}(A, \mathit{ec}1\_1\_1).$
   A homotypic interaction occurs if the corresponding protein contains the EC number 1.1.1 (with NAD or NADP as acceptor).
5. $\mathit{high\_reliability\_ppi}(A, A) :- \mathit{subcellular\_location}(A, \mathit{cytoplasmic}).$
   A homotypic interaction occurs if the corresponding protein is located in the cytoplasmic of the cell.
6. $\mathit{high\_reliability\_ppi}(A, B) :- \mathit{correlation}(A, B, C), \mathit{gteq}(C, 0.806).$
   Protein A interacts with protein B if their gene expression correlation coefficient is greater than 0.806.
7. $\mathit{high\_reliability\_ppi}(A, A) :- \mathit{correlation}(A, B, C), \mathit{lteq}(C, -0.043).$
   A homotypic interaction corresponding to protein A occurs if there exists a protein B that the gene expression correlation coefficient between A and B is less than -0.043.
8. $\mathit{high\_reliability\_ppi}(A, B) :- \mathit{dr\_go}(A, \mathit{go}0000778), \mathit{correlation}(A, B, C), \mathit{gteq}(C, 0.297).$
   Protein A interacts with protein B if A is related to the GO term "C:condensed nuclear chromosome kinetochore" and there exists a protein B that the gene expression correlation coefficient between A and B is greater than 0.297.

Fig. 2 ILP rules obtained.

Having provided training examples and background knowledge described above, Figure 2 shows some rules obtained as well as description for each rule.

We also conducted ten-fold cross-validation to examine the accuracy of the learner, by varying k as in Table 1).

Table 1 Accuracy obtained by varying IST hit threshold k

IST hit threshold k	Accuracy
1	64.28 ± 2.58 (%)
2	79.84 ± 2.34 (%)
3	87.03 ± 1.36 (%)

5. Conclusions

In this paper, we have applied ILP to the problem of predicting protein-protein interactions with high reliability. We have used two kinds of background knowledge by extracting useful information from protein databases, and by using a computational approach. The results obtained are promising, both for the comprehensibility of rules generated using ILP and the cross-validated accuracy. In future work, we are attempting to exploit more background knowledge, such as by using homology search. The proposed approach also needs to be tested on other data sets, for example BIND[1], DIP[27], and MIPS[5].

Acknowledgements

This research is supported by the grant-in-aid for scientific research on priority areas (C) "Genome Information Science" from the Japanese Ministry of Education, Culture, Sports, Science and Technology.

References

1) G. D. Bader, I. Donaldson, C. Wolting, B. F. F. Ouellette, T. Pawson, and C. W. V. Hogue.

Bind – the biomolecular interaction network database. *Nucleic Acids Research*, 29:242–245, 2001.

2) A. Bauer and B. Kuster. Affinity purification-mass spectrometry: Powerful tools for the characterization of protein complexes. *Eur. J. Biochem.*, 270(4):570–578, 2003.

3) J. R. Bock and D. A. Gough. Predicting protein-protein interactions from primary structure. *Bioinformatics*, 17(5):455–460, 2001.

4) T. Dandekar, B. Snel, M. Huynen, and P. Bork. Conservation of gene order: A fingerprint of proteins that physically interact. *Trends Biochem. Sci.*, 23(9):324–328, 1998.

5) Comprehensive Yeast Genome Database. http://mips.gsf.de/genre/proj/yeast/index.jsp.

6) SWISS-PROT database. http://www.expasy.ch/sprot.

7) Yeast Interacting Proteins Database. http://genome.c.kanazawa-u.ac.jp/Y2H/.

8) InterPro database concerning protein families and domains. http://www.ebi.ac.uk/interpro/.

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

10) M. Deng, F. Sun, and T. Chen. Assessment of the reliability of protein-protein interactions and protein function prediction. In *Pacific Symposium on Biocomputing*, pages 140–151, 2003.

11) Protein families database of alignments and HMMs. http://www.sanger.ac.uk/Software/Pfam/.

12) C. S. Goh, A. A. Bogan, and M. Joachimiak et al. Co-evolution of proteins with their interaction partners. *J. Mol. Biol.*, 299(2):283–293, 2000.

13) A. Grigoriev. A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage t7 and the yeast *saccharomyces cerevisiae*. *Nucleic Acids Res.*, 29(17):3513–3519, 2001.

14) T.Ito, T.Chiba, R.Ozawa, M.Yoshida, M.Hattori, and Y. Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. In *Proc. Natl. Acad. Sci. USA 98*, pages 4569–4574, 2001.

15) L. Lu, H. Lu, and J. Skolnick. Multiprospector: An algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins*, 49(3):350–364, 2002.

16) E.M. Marcotte, M.Pellegrini, and H.L.Ng etal. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285(5428):751–753, 1999.

17) E. M. Marcotte, I. Xenarios, and D. Eisenberg. Mining literature for protein-protein interactions. *Bioinformatics*, 17(4):359–363, 2001.

18) L. R. Matthews, P. Vaglio, and J. Reboul 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.

19) S. Muggleton. Inverse entailment and progol. *New Generation Computing*, 13:245–286, 1995.

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

21) S. K. Ng and S. H. Tan. Discovering protein-protein interactions. *Journal of Bioinformatics and Computational Biology*, 1(4):711–741, 2003.

22) PROSITE: Database of protein families and domains. http://kr.expasy.org/prosite/.

23) Gene Ontology. http://www.geneontology.org/.

24) T. Oyama, K. Kitano, K. Satou, and T. Ito. Extracting of knowledge on protein-protein interaction by association rule discovery. *Bioinformatics*, 18(5):705–714, 2002.

25) M.Pellegrini, E.M. Marcotte, and M.J.Thompson et al. Assining protein functions by comparative genome analysis: Protein phylogenetic profiles. In *Proc. Natl. Acad. Sci. USA 96(8)*, pages 4285–4288, 1999.

26) Protein Information Resource. http://pir.georgetown.edu/.

27) L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J.U. Bowie, and D. Eisenberg. Dip: The database of interacting proteins: 2004 update. *Nucleic Acids Research*, 32:449–451, 2004.

28) T. Sekimizu, H. S. Park, and J. Tsujii. Identifying the interaction between genes and gene products based on frequently seen verbs in medline abstracts. *Genome Informatics*, pages 62–71, 1998.

29) G. P. Smith. Filamentous fusion phage: Novel expression vectors that display cloned antigens on the virion surface. *Science*, 228(4705):1315–1317, 1985.

30) P. T. Spellman, G. Sherlock, and M. Q. Zhang et al. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, 9(12):3273–3297, 1998.