An Ontology-Empowered Model for Annotating Protein-Protein Interaction Data: a Case Study for Budding Yeast

Arash Shaban-Nejad and Volker Haarslev
Dept. Computer Science and Software Eng.,Concordia University, Montreal, Quebec, H3H2J8, Canada, Tel: (514) 848-2424 (Ext. 7122)
{arash_sh, haarselv}@cs.concordia.ca

Abstract

This paper reports on our experience in modeling and employing ontology-inferred knowledge to support and improve data mining tasks of yeast protein interactions for knowledge discovery. This objective has been accomplished by providing simplified access to units of intersecting proteome data and information from different biological databases and bio-ontologies, and utilizing a logical framework to answer questions from biologists.

1. Introduction

Proteins are crucial in biological systems. Most protein functions depend on interactions with other molecules. In addition, protein–protein interactions provide rich information on the fundamental aspects of cellular life, and can be used in areas like target selection in drug discovery. To study the functional interactions of different proteins, one needs to access to a consistent set of data about protein interactions. These data are scattered over various databases and model organisms. Acquisition, representation, integration, and validation of so much stored scientific data of various types need a combination of machine learning and knowledge representation methods, including semantic interpretation, structural and functional similarity assessment, control relationships, properties, and various annotation and validation techniques. Many of these techniques require a formal description of part of an intended domain in the real world. Ontologies provide a set of shared and precisely defined terms in various degrees of formality to describe a particular domain of interest.

This paper reports on our experience in extracting knowledge from current data and information sources of the yeast protein interactions, to improve protein interaction data mining and text mining. We have achieved this purpose by modeling an integrated logic-based ontology to represent knowledge and answer biological questions. In our approach, we have used graph-based data mining algorithm [4]. Also for the purpose of our study, we have considered mutually interacting protein pairs in the budding yeast *Saccharomyces cerevisiae*, which is one of the best studied of all organisms with a rich amount of available data and knowledge in cell biology and genetics [1]. Most yeast proteins can be connected in a large network of interactions. Yeast has a very dynamic protein interaction network, with more than 30,000 interactions [2].

Many scientists in labs all around the world generate a large amount of protein data using several applications. These data are highly volatile, complex, inter-related, and heterogeneous. They have different types, algorithms (BLAST, FASTA, pSW), forms, and implementations (WU- BLAST, NCBI-BLAST); they are generated by various communities and service providers (NCBI, EBI, DDBJ) [3]. Protein complexes sometimes have different behaviors from their basic elements. Even one domain does not always fulfill the same functions. In addition, different synonyms, IDs/accession numbers, relations, interactions, and functions as free text descriptions cause more confusion. Ontologies provide a shared understanding from these heterogeneous data and information by defining axioms, concepts and properties. Here we present the usability of an integrated ontology-based framework in a graph-based data mining system dedicated to the collection, validation, and integration of protein interaction data.

2. Protein Interactions

2.1. Interactions with other molecules

Most protein functions depend on interactions with other molecules, such as the nucleic acids DNA or RNA, solvent molecules (e.g., water), and metal ions [5].

2.2. Protein-protein interaction

Interactions of proteins with other proteins provide precious information about their functions and biological roles with directions for phenotypic examination of mutants for the novel genes [1]. In addition, this information can aid in the discovery of a number of protein-binding domains or motifs, used in diverse signaling pathways [6].

2.3. Methods for interaction analysis and their data

Most current knowledge of yeast protein interactions comes from a few methodologies, including the yeast two-hybrid assay, the purification of protein complexes, and their analysis by mass spectrometry [2]. To study all possible protein-protein interactions, biologists often use the two-hybrid screening system where ″the interacting proteins are decoded by sequence tagging of the plasmid inserts″ [1].

2.3.1. Two-hybrid assay interactions. The two-hybrid system uses two hybrid proteins that reconstitute a transcription factor (TF) when they interact. This TF can switch on a reporter gene. About 7000 two-hybrid protein interactions are available in databases, derived from small and large screens [2].

2.3.2. Purification and analysis by mass spectrometry. Proteins interact within stable protein complexes. The components of these complexes can be identified by complex purification. An approach called tandem-affinity purification (TAP) along with mass spectrometry [39] can characterize protein complexes in *Saccharomyces cerevisiae* [41].

Comparing interaction datasets in [2] shows that various methods produce different results. Biologists need to have a consistent view of these datasets to be aware of these differences.

3. Integrating protein interactions data with yeast proteome datasets

In order to integrate the yeast protein-protein interactions datasets with other data scattered over distributed data sources, one need to browse the relationship between gene expression and protein interactions first [20] to allow evaluating “interaction datasets using large-scale gene expression profiling as benchmark” [2]. Integration of various interactions datasets remains a difficult challenge for bioinformatics, especially when interactions have to be done with expression data, homology, structures, or localization in a cell [11].

Protein interactions can be explained based on their actions on other proteins (similar to the interactions between enzymes and their substrates [11]). Different parameters describing protein interactions such as Concentration, Localization, Cleavage, Binding site, Covalent and Non-covalent modifications are described in [11].

3.1 Major information sources

Many datasets and information resources contribute to provide sufficient data and information to describe protein interactions. To reduce the complexity, in this study, we have focused on static protein interactions, which deal with less parameter in compare with the dynamic interactions [11]. The datasets and information about protein-protein interactions are scattered over various data sources [12, 11]. For knowledge discovery, one needs to search in distributed databases, literatures, and existing bio-ontologies. Some of our major resources are:

- DIP (Database of Interacting Protein) [13]
- SGD (Saccharomyces Genome Database)
- BIND (Biomolecular Interaction Network Database) [10]
- Yeast protein-protein interactions database [14]
- The Protein-Protein Interaction Server [15]

In addition, the following bio-ontologies are used as other information resources: Protein Ontology (PO) [16], Gene Ontology (GO) [17], FungalWeb Ontology (FWO) [18], and TRANSFAC Ontology [19].

As mentioned, one of the major resource in our project is BIND [10], which curates and archives physical interactions between bio-molecules from the literature, using a standard data representation. BIND is a suitable source for developing interaction networks into pathways. It is available in different format, such as ASN.1 (text), XML, and Flat File. In our project,

we have adapted the available XML version of BIND witch a similar structure to OWL format.

The Protein Ontology [16] provides common terminologies and classification for capturing knowledge about protein domain.

The "TRANSFAC" [19] is a collection of databases that deal with information about gene expression.

3.2. The integrated ontology structure

We considered the FungalWeb Ontology [18], which is implemented in OWL-DL as the basis of our integrated ontology. The FungalWeb is a resource for fungus and enzyme-related terminologies and concepts, which mostly come from the NCBI taxonomy database [22], NEWT [21], BRENDA [23], SwissProt [24], and some commercial enzyme vendors. This ontology also reuses existing domain-specific bio-ontologies such as Gene Ontology (GO) and TAMBIS [25].

Kingdom	Fungi
Phylum	*Ascomycota*
Class	*Hemiascomycetes*
Order	*Saccharomycetales*
Family	*Saccharomycetaceae*
Genus	*Saccharomyces*
Species	*Sacchromyces. cerevisiae*

Fig 1. Sacchromyces. Cerevisiae (budding yeast) classification in the FungalWeb Ontology

One can find the attributes and the corresponding enzymes for the fungus organism “budding yeast” in the FungalWeb ontology. Other vocabularies related to the corresponding proteins for each fungus and their interactions come from the databases and bio-ontologies mentioned in Section 3.1.

To have a coherent knowledge-based system that combines and infers implicit knowledge from different databases and ontologies in a domain of interest, the mentioned databases and ontologies should be linked to each other (Figure 2). By integrating and reusing existing databases and ontologies, the integrated system would be more efficient for information retrieval, knowledge discovery, query answering, and decision-making [26].

We tried to map the databases’ elements to a set of ontological concepts, and we have reused vocabularies from generic bio-ontologies in a unified ontological framework. To reuse these ontologies together, they have been combined and merged into a new ontology. For this purpose, the ontologies were brought into mutual agreement. In some cases, the integration has been accomplished by relating analogous concepts or relationships from distributed resources to each other by an equivalence relationship. The integration is done at two levels: Data integration (normalizing various formats of data that are extracted from different sources) and Semantic Integration (specifying the applicable datasets with their semantic relationships).

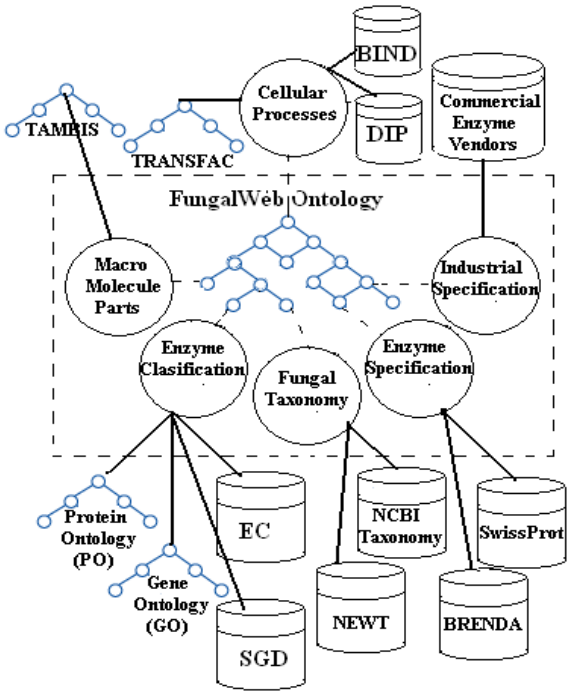

Fig 2. The structure of the integrated ontology for protein interaction including available domain related databases and bio-ontologies

After finding the regions of overlap in the ontologies, the concepts with close semantics are aligned in consist and coherent way.

We have used PROMPT [27] as a Protégé plug-in for automated ontology merging and alignment. PROMPT works based on similarity matching which sometimes generate imprecise results. So, to control the quality of the results, human supervision was needed at this stage.

3.3. Integrated ontology-driven data mining Framework

Data mining is defined as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data" [38]. Graphs are currently used to model protein structures for the identification of active site clusters, folding clusters, and aromatic clusters, in relation to thermodynamic stability, and the analysis of protein-protein interactions. (For more information applying graph theory to protein interactions see: [9].)

As stated in [34] “a molecular interaction networks can be mapped to labeled graphs. Every node of the graph represents a distinct amino acid residue in a protein and has the residue type as its label”. Different graph representations have been proposed [34], ranging from coarse (each node is a secondary structure segment) [35] to fine (each node is an atom) [36] representations. A part of a graph where each node is attached to the others with at least k edges is called k-core. As stated at [33] highest k-core is a “central most densely connected region of a graph, which may represent molecular complexes”.

BIND [10] uses graph theory to represent molecular interactions. For the graph annotation, BIND employs graph coloring techniques connected with the Gene Ontology. Using graph theory, BIND has discovered 7000 yeast interactions among 3000 proteins.

Some of the major tasks in data mining are finding patterns (i.e., association rules), classification, and grouping (or clustering) [7]. A vast amount of valuable background knowledge is distributed in various databases, texts (e.g., publications), or database annotations. Current available protein interaction databases and information resources are incomplete, incoherent and in some cases inconsistent. Today communities do not use a standard nomenclature for genes/proteins, which makes proteomic data mining difficult and ineffective [8]. Ontologies can provide a shared vocabulary for existing proteomic sources. Ontology can utilize multi-relational data and can improve the efficiency of large-scale data mining problems. We have employed ontologies along with data mining methods to design an integrated system for analyzing protein interactions. Particularly we are applying data mining algorithms to analyze the yeast protein interaction dataset. By using data mining methods, we survey and discover patterns and differences at various biological resolutions for a given concept.

4. Logic-based representation of biological knowledge

The FungalWeb Ontology is implemented in OWL-DL format, giving it the advantages of description logics (DL) to support maximum expressivity without losing computational completeness. DL describes knowledge in terms of concepts and relations that are used to automatically derive classification taxonomies. Concepts in description logics are defined in terms of descriptions using other roles and concepts. In this way, the DL reasoner (such as RACER [28]) can automatically classify Enzyme as a kind of Protein [25]. The semantic is not clear in graph-based representation, which is popular in current protein interaction data mining systems, so we tried to solve this problem by using description logic representation.

Description Logics have clear semantics, so it is possible for the available reasoner (in our case, RACER) to verify ontological consistency and coherency.

There are certain ways to convert graph representation to logic [29]. This conversion is done manually in our project. Description logics are introduced as a language with transitive closure on roles [30]. Languages based on description logics, such as OWL-DL, can be related to directed acyclic graphs through sets of nodes similar to concepts, roles similar to graphs, and transitive closure of roles similar to reachability [31]. Therefore, techniques in description logics, like the Tableau-Based Decision Procedure, can lead to advances in graph theory, and vice versa [31].

5. Evaluation and Querying

In our approach, the evaluation is pragmatic, accomplished by assessing the ontology to satisfy the requirements of our application. The DL reasoner (RACER) assists in maintaining semantic consistency. Although we validate the biological data and relations by citing their origin (database or literature), validation by the domain expert is also necessary. We use RACER as a DL reasoning system, which supports T-Box (axioms about class definitions) and A-Box (assertions about individuals), for reasoning in the integrated ontology and checking the A-Box and T-Box consistency. RACER solves the posed subsumption problems very quickly.

We use nRQL (new RACER Query Language) [32] as a query language based on RACER. The querying can be done by experts or by a software agent. It is possible for the answering agent to use automated reasoning methods when deriving answers to queries where the knowledge necessary may be found in multiple knowledge-bases. In the current state of our integrated ontology, the knowledge source for querying is determined manually. Some queries in the system that might be interesting for biologists studying protein-protein interactions can be defined as follows:

1. Assume there is an interaction between two proteins X and Y. Protein Z is homologous to X, and protein W is homologous to Y. Do Z and W interact?
2. What is the corresponding organism for a protein?
3. Are the two proteins from the same compartment?
4. Was the interaction saturable?
5. Are the two proteins involved in an interaction known to be involved in the same process?

Queries 2 and 3 can be studied from the yeast two-hybrid method, and queries 4 and 5 can be studied from the affinity purification method [33].

6. Challenges and Future works

Knowledge discovery process includes different phases, such as data preparation, integration, and transformation. Each of these phases can benefit from an ontology. We integrated several protein interaction resources in a logic-based ontological framework, which can be used as a knowledge-base for querying and capturing annotation data in the protein interaction domain. By using the ontology with the graph-based data mining algorithm, we would be capable of doing some knowledge discovery in the domain through designed queries.

One of our major issues was semantic integration. The protein interaction data must be linked to complementary biological information, but these data and information are inconsistent, complex, and highly volatile. At the ontology integration stage, we were faced with the problem of mismatching between ontologies in language and model levels. There were differences in conceptualization and in the way the conceptualization is specified. Future work will focus on improving ontological structure and considering relations and protein interactions with small molecules. The interactions of small molecule with proteins play an important role for determining the effects of drugs (which are mostly small molecules) [40] in the human body.

7. References

[1] T. Ito, K. Tashiro, S. Muta, R. Ozawa, T. Chiba, M. Nishizawa, K. Yamamoto, S. Kuhara, Y. Sakaki, "Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins" *Natl Acad Sci* USA. 2000, 1; 97(3), pp. 1143-1147.

[2] P. Uetz, and A. Grigoriev, "The yeast interactome," Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics (EGGPB), J.Wiley and Sons, 2005.

[3] K. Kumpf, J. Fluck, and M. Hofmann, "A Protein Ontology from Large-scale Textmining?" presented at Protégé Workshop, Manchester, Sep 2003.

[4] D.J. Cook, and L.B. Holder, "Graph-Based Data Mining, " IEEE Intelligent Systems, 2000: 15 (2), pp. 32-41.

[5] J. Mitchell, "Protein Interactions (I)". Cambridge, 2003. http://www-mitchell.ch.cam.ac.uk/courses/int1.html

[6] T. Pawson, "Protein modules in signal transduction." Springer; New York, 1998.

[7] I.K. Ravichandra Rao, "Data Mining and Clustering Techniques," presented at DRTC Workshop on Semantic Web, DRTC, Bangalore, 8-10 Dec. 2003.

[8] R.A. Saavedra, R. Baughman, D. Tagle, and R. Stewart, "Protein-Protein Interaction Maps for the Mammalian Nervous System," presented at Workshop on Protein - Protein Interaction Maps for the Mammalian Nervous System, Maryland, Nov. 2004.

[9] S. Vishveshwara, K.V. Brinda, and N. Kannan, "Protein structure: Insights from graph theory," *J. of Theo. and Comp. Chem.*, 2002, 1(1), pp. 187–211.

[10] G.D. Bade, C.W. Hogue, "BIND-a data specification for storing and describing bimolecular interactions, molecular complexes and pathways," *Bioinformatics,* 2000 ,16(5), pp. 465-77.

[11] P. Uetz, T. Ideker, and B. Schwikowski, "Visualization and integration of protein-protein interactions." In *Protein-Protein Interactions – A Molecular Cloning* Cold Spring Harbor Laboratory Press 2002.

[12] Rosalind Franklin Center for Genomics Research, "Protein interaction databases". http://www.hgmp.mrc.ac.uk/GenomeWeb/prot-interaction.html

[13] I. Gilfillan, "A database of proteins that are known to Interact," *Genome Biology* 2000, 1:reports220.

[14] Yeast protein-protein interactions database http://www.hsls.pitt.edu/guides/genetics/tools/protein/interaction /URL1048784666/info

[15] The Protein-Protein Interaction Server http://www.biochem.ucl.ac.uk/bsm/PP/server/

[16] A.S. Sidhu, T.S. Dillon, E. Chang, "Creating a Protein Ontology Resource," *IEEE Computational Systems Bioinformatics Conf.* – Workshops (CSBW'05), 2005, pp. 220-221.

[17] M.A. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, and etc., "The Gene Ontology (GO) database and informatics resource.", Nucleic *Acids Res*. 2004; 32, pp. 258–261.

[18] A. Shaban-Nejad, C.J.O. Baker, V. Haarslev, and G. Butler, "The FungalWeb Ontology: Semantic Web Challenges in Bioinformatics and Genomics", In the proceedings of *4th Intl. Semantic Web Conference (ISWC)*, Galway, Ireland. *LNCS,* Vol. 3729, Nov. 2005, pp. 1063-1066.

[19] E. Wingender, "TRANSFAC, TRANSPATH and CYTOMER as starting points for an ontology of regulatory networks*" In Silico Biol*. 2004;4(1), pp. 55-61.

[20] 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*, 2001, 29, pp. 3513-3519.

[21] I.Q.H. Phan, S.F. Pilbout, W. Fleischmann, A. Bairoch, "NEWT, a new taxonomy portal", *Nucleic Acids Research*; 2003, 31(13), pp. 3822-3823.

[22] D.L. Wheeler, C. Chappey, A.E. Lash, D.D. Leipe, T.L. Madden, G.D. Schuler, T.A. Tatusova, and B.A. Rapp, "Database resources of the National Center for Biotechnology Information.", *Nucleic Acids Research; 2000,* 28(1). pp. 10-14.

[23] I. Schomburg, A. Chang, C. Ebeling, M. Gremse, C. Heldt, G. Huhn, and D. Schomburg, "BRENDA, the enzyme database: updates and major new developments.", *Nucleic Acids Research; 2004,* 32

[24] A. Bairoch, "The ENZYME database in 2000". *Nucleic Acids Research;2000,* 28, pp. 304-305.

[25] P.G. Baker, A. Brass, S. Bechhofer, C. Goble, N. Paton, and R. Stevens, "TAMBIS- Transparent Access to Multiple Bioinformatics Information Sources." *Proc Int Conf Intell Syst Mol Biol;1998,* 6, pp. 25-34.

[26] M. Uschold, M. Healy, K. Williamson, P. Clark, and S. Woods. "Ontology reuse and application." Presented at Formal Ontology in Information Systems (*FOIS'98)*, Trento, Italy, JUNE 6-8, 1998.

[27] PROMPT: a tool for managing multiple ontologies in Protégé. Available at: http://protege.stanford.edu/plugins/prompt/prompt.html

[28] V. Haarslev, and R. Möller, "RACER System Description." In the Proc. of *Int.l. Joint Conference on Automated Reasoning, IJCAR'2001*, Siena, Italy, Springer-Verlag, Berlin, June 18-23, 2001, pp. 701-705.

[29] A.M. Rassinoux, R.H. Baud, C. Lovis, J.C. Wagner, J.R. Scherrer, "Tuning up conceptual graph representation for multilingual natural language processing" *in medicine," In Proc. of 6th Intl. Conference on Conceptual Structures, Conceptual Structures: (ICCS'98),* Montpellier, France, Aug. 1998, pp.390-397.

[30] F. Baader, "Augmenting concept languages by transitive closure of roles: An alternative to terminological cycles.", In *Proc. of the 12th Intl. Joint Conference on AI*, 1991, pp. 446-451.

[31] D. Cantone, D. Calogero, and G. Zarba, "A Tableau-Based Decision Procedure for a Fragment of Graph Theory Involving Reachability and Acyclicity," *in proc. Of the Intl. Conference on Automated Reasoning with Analytic Tableaux and Related Methods (TABLEAUX 2005)*, Koblenz, Germany, 2005, pp. 93-107

[32] M. Wessel, R. Möller,"A High Performance Semantic Web Query Answering Engine.", *Int. Workshop on DL.* (*DL2005*), 2005, Edinburgh, Scotland.

[33] M. Dumontier, "Protein interactions." Presented at *Proteomics workshop 2005 in* Canadian bioinformatics workshop series, Montreal, 2005.

[34] J. Huan, D. Bandyopadhyay, W. Wang, J. Snoeyink, J. Prins, and A. Tropsha, "Comparing Graph Representations of Protein Structure for Mining Family-Specific Residue-Based Packing Motifs." *Journal of Computational Biology.* Jul 2005, Vol. 12, No. 6, pp. 657-671

[35] H. Grindley, P. Artymiuk, D. Rice, and P. Willet, "Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm," *J. Mol. biol.,* 1993, 229, pp. 707–721.

[36] D. Jacobs, A. Rader, L. Kuhn, and M. Thorpe, "Graph theory predictions of protein flexibility," *Proteins: Struct.Funct. Genet.,* 2000: 44, pp.150-155.

[37] S. Sundararaj, "Protein Pathways and Pathway Databases," Presented at *Proteomics workshop 2005 in* Canadian bioinformatics workshop series, Montreal, 2005.

[38] D.J. Cook, L.B. Holder, "Graph-Based Data Mining." IEEE Intelligent Systems. 2000: 15 (2), pp. 32-41.

[39] A. Bauer, and B. Kuster, "Affinity purification-mass spectrometry: Powerful tools for the characterization of protein complexes," Eur. J. Biochem. 270, 570–578 (Feb 2003)

[40] H. Feldman, "Small Molecule-Protein Interactions" Presented at *Proteomics workshop 2005 in* Canadian bioinformatics workshop series, Montreal, 2005

[41] A.C. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, and etc. "Functional organization of the yeast proteome by systematic analysis of protein complexes." *Nature 2002*, 415, 141-147.