Using ontologies for querying and analysing protein-protein interaction data

Mario Cannataro, Pietro Hiram Guzzi

Bioinformatics Laboratory, Department of Experimental Medicine and Clinic
University Magna Graecia, Catanzaro, Italy
{cannataro,hguzzi}@unicz.it

1 Introduction

Many wet lab experiments lead to the accumulation of a large amount of data related to interaction among proteins [1] also referred as Protein to Protein Interaction (PPI) data. The whole set of protein interactions of a single organism is also referred to as Protein to protein Interaction Network (PIN) and it is built from a set of binary interactions. PINs have been easily modeled by using undirected graphs [2] where nodes are associated to proteins, and edges represent interactions among proteins. The high dimension of this graph makes infeasible the manual inspection even for simple organisms, so the study of PINs requires graph-based computational methods.

PPI databases, such as the Database of Interacting Proteins (DIP) [3], are often publicly available on the Internet and offer to the user the possibility to retrieve data of interest through simple querying interfaces. User, in fact, can conduct a search by the insertion of: (i) one or more protein identifiers, or (ii) a protein sequence. Results may consist of, respectively, a list of proteins that interact directly with the seed protein or that are at distance k from the seed protein in the PIN. It is impossible to formulate even simple queries involving biological concepts or annotations, such as: *retrieve all the interactions that are related to glucose synthesis*‘. Nevertheless, these annotations there still exist and are spread in different data sources. The main hypothesis of this paper is that annotating PPI data with biological information may result in more rich querying interfaces and in more powerful PINs analysis algorithms that may use such biological information [4]. This work presents a first prototype of a system system able to adding to actual data the information coming from ontologies such as Gene Ontology [5] and from other sources. Moreover the proposed system allows the use of annotated data into an analysis pipeline.

The annotation of a PPI network consists of three main phases: (i) retrieval of PPI data (**Data Extraction Module**), (ii) retrieval of existing annotations for that data (**Metadata Extraction Module**), (iii) generation of annotated interactions and storage into the annotated database. Initially, the proposed system queries the existing interaction database and retrieves data about interactions. Then the protein identifiers are used to find related annotations. For instance, the Gene Ontology Annotation Database [6] (GOA) can be queried by

using the UniProt identifiers or Gene Ontology terms. Finally, data are merged together and encoded by using an XML-based syntax, and stored into the annotated database. Figure 1 depicts the architecture of the system for extracting annotations from Gene Ontology and for annotating the PPI database.

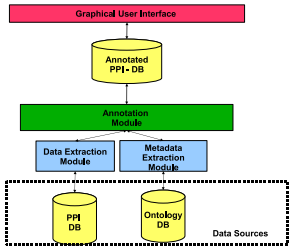

Fig. 1. The Architecture for annotating PPI databases

Main advantage of such system is the possibility to retrieve interactions, non only proteins whose nodes have a given annotation. Let us consider protein MEC1 of yeast and its interacting partners. Let us consider, moreover, the *kinase activity* process. When a user searches in existing databases it will retrieve the interactions: (MEC1, TEL1), (MEC1, RNR1). Successively, he/she has to check the annotation manually to discover which proteins are annotated with kinase activity. By using the annotated PPI database user can directly specify the process retrieving desired informations.

Such a system could be useful not only for the semantic search of data, but also for the semantic-based analysis of PPI data. The analysis of PPI networks is usually done by using graph-based algorithms, and associating graph properties to biological properties of the modeled PPI. The availability of annotated data could enable the development of novel algoritms able to gather such information.

References

1. Uetz, a.: A comprehensive analysis of proteinprotein interactions in saccharomyces cerevisiae. Nature **403** (2000) 623627
2. West, D.B.: Introduction to Graph Theory. Prentice Hall, NY (August 2000)
3. Salwinski, S.e.a.: The Database of Interacting Proteins: 2004 update. Nucl. Acids Res. **32**(suppl1) (2004) D449–451
4. Cannataro, M., Guzzi, P.H., Veltri, P.: Using ontologies for annotating and retrieving protein-protein interactions data. In: Computer-Based Medical Systems, 2009. CBMS 2009. 22nd IEEE International Symposium on. (Aug. 2009) 1–5
5. Harris, M.A., et al: The gene ontology (go) database and informatics resource. Nucleic Acids Res Nucleic Acids Res **32**(Database issue) (January 2004) 258–61
6. Camon, E., et al: The gene ontology annotation (goa) database: sharing knowledge in uniprot with gene ontology. Nucleic Acids Res **32**(Database issue) (January 2004)