

e-Science Tools for the Analysis of Complex Systems

Olusola C. Idowu, Steven J. Lynden and Peter Andras

School of Computing Science, University of Newcastle upon Tyne, Newcastle upon Tyne, NE1 7RU, UK

Abstract

Many real world complex systems (e.g., protein-protein interaction networks in living cells, the Internet) can be conceptualised as graphs of interacting components, where the components are the nodes and the interactions are the edges of the graph. Identifying the functionally important components of complex network systems can be of considerable interest in many real world applications. Consequently, analysing graphs of complex networks to detect their vulnerable components and to establish their structural integrity is very important. The e-Science Solutions for the Analysis of Complex Systems (eXSys) project aims to develop e-Science analysis tools for analysing vulnerabilities in complex networks in particular the protein-protein interaction network. In this paper we describe the techniques we used for collecting protein-protein interaction data for numerous organisms from heterogeneous bioinformatics databases and the methods used for data analysing.

1. Introduction

A common feature of many complex systems is that they can be modelled as networks of interacting components. Figure 1.1 illustrates how intracellular protein interactions can be modelled in this way. It has recently been discovered that many complex networks that exist in the world around us (including the Internet and networks of biochemical reactions in living cells) [2] are far from random and that they usually possess a small number of nodes that are unusually important to the integrity of a network as a whole. As a result of this, such networks are vulnerable to targeted interventions which aim to damage a network by inhibiting the functionality of important nodes.

Within living cells, individual proteins bind to each other to form complexes that perform various functions that contribute towards keeping cells alive. Complex protein interaction networks consist of thousands of proteins, where an interaction between two proteins exists if they are able to bind with each other. When the human body is invaded by a microbial infection we may fight the invasion with drugs, such as antibiotics, which aim to destroy the pathogen organism causing the infection while minimising any side effects. Discovering a protein that is vital to the network integrity of a pathogen protein interaction network, but at the same time not vital within human cells, potentially identifies a target protein which drugs can be designed to inhibit. By constructing models of protein interaction

networks and performing graph theoretic and statistical analysis, proteins which are highly important towards the survival of an organism may be discovered which can potentially be exploited as drug targets.

2. Data integration and database system architecture

2.1 Protein interaction data sources

The data required to construct models of protein interaction networks may be obtained from multiple distributed heterogeneous data sources. A data management tool has been developed which is capable of automatically gathering data from various sources on the World Wide Web and transforming this data into a common data format. This tool has been used to construct a single comprehensive data source of protein interactions by integrating data from the Database of Interacting Proteins (DIP) [3], the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) [4] and the Kyoto Encyclopaedia of Genes and Genomes (KEGG) Pathway database [5]. The DIP contains experimentally detected protein interaction data obtained from various sources including journal articles and other protein interaction databases. The DIP currently contains approximately 20,000 protein interactions among roughly 100 organisms.

The KEGG consists of a variety of databases containing proteomics and genomics information compiled from a vast amount of empirical data. The KEGG Pathway database

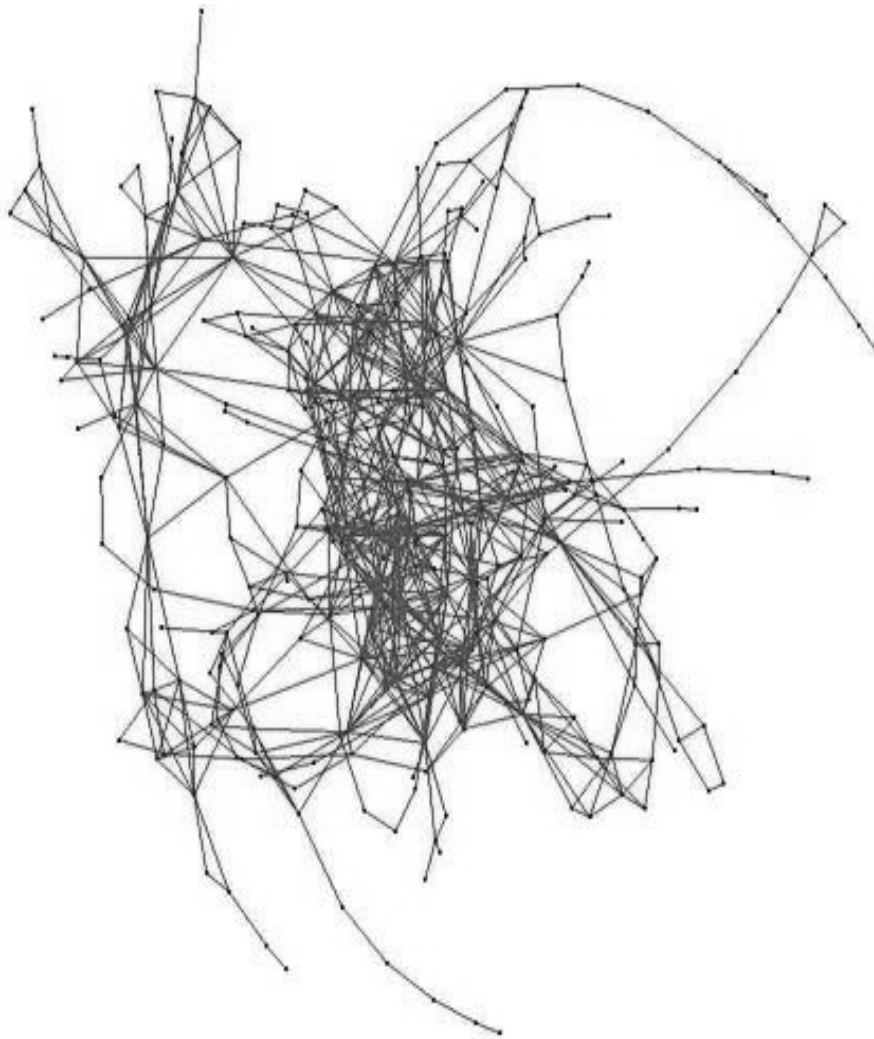


Figure 1.1: Protein interaction network of *Staphylococcus aureus*

The figure illustrates a protein interaction network constructed using protein interaction data for Staphylococcus aureus, a single-celled bacterium. Although many proteins may only be present within a cell only at certain specific times, it is possible to construct a static representation such as the one illustrated by considering all the interactions between proteins that can possibly take place within a cell. Analysing network topologies such as this enables the identification of proteins that are of high relative importance to the integrity of the network.

contains information concerning metabolic pathways, regulatory pathways and molecular complexes. Metabolic pathways are arranged into categories, such as carbohydrate metabolism, energy metabolism, and nucleotide metabolism etc. Each of these categories is subdivided into a number of subcategories for which there are individual reference pathway

maps. The reference pathway maps display known relationships between gene products, and each individual reference pathway map may be retrieved as an organism specific pathway map containing gene products and relationships that apply to specific organisms only. Among the relationships are enzyme-enzyme relations where two enzymes catalyze successive

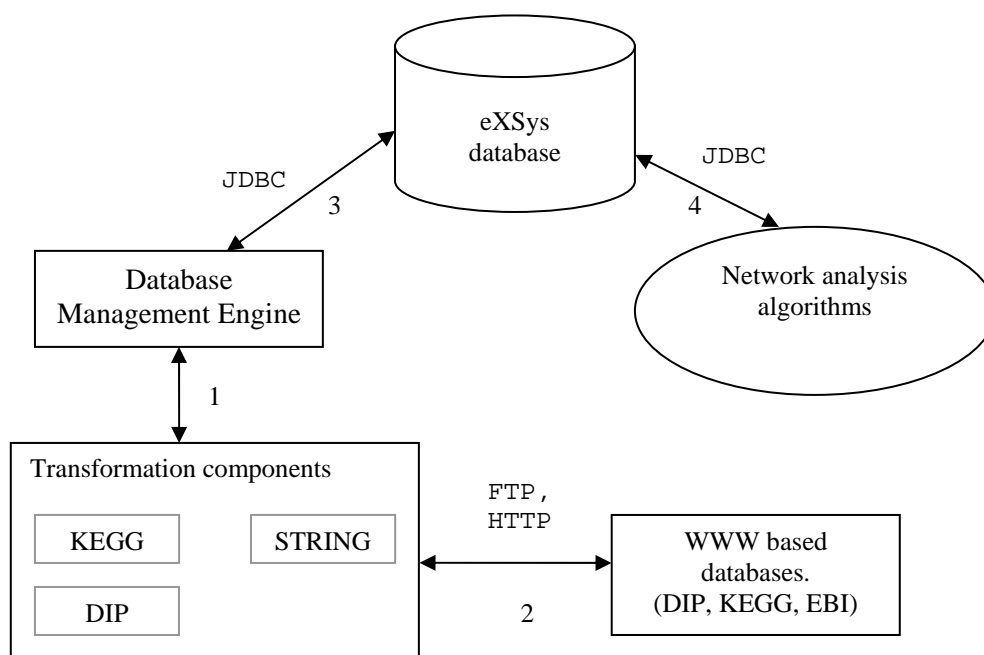


Figure 2.1: Data integration process and software architecture.

Data integration is managed by the database management engine, an autonomous component that utilises the data source specific transformation components (1) to retrieve and then transform data (2) into the format used by the eXSys database. Once data has been transformed, it is integrated into the eXSys database (3). This data may then be used by the various network analysis algorithms (4). Almost all proteomics databases, including the four sources currently utilised, are updated frequently and the database management engine periodically repeats this process to add any new data that has been made available from each data source. The transformation components and data management engine are implemented using Java. We used Java Database Connectivity (JDBC) is utilised for communication between the database, data management engine and network analysis algorithms.

reactions in a metabolic pathway. It is from these enzyme-enzyme relations that we derive protein interactions.

In addition to experimentally detected interactions between proteins, protein interactions may also be predicted using information such as protein structure and genetic information. The STRING database contains predicted functional associations between proteins, derived from analysing genomic associations between the genes which encode them, as described in [4]. Specifically, interactions between clusters of orthologous groups of genes are predicted, and each interaction between two groups is assigned a confidence score estimating the precision of the predicted interaction. Orthologous groups contain genes belonging to different species, which have evolved from a common ancestral

gene and often perform similar functions. Individual proteins are mapped to orthologous groups, allowing a protein interaction to be predicted by detecting the presence of two proteins, present in the same organism but different orthologous groups, where the groups are predicted to interact.

2.2 Data Integration and Management

The requirements of the data integration process are to extract from existing data sources the data required to perform protein interaction network analyses and store this data in the local eXSys database (a MySQL relational database). The database is populated by transforming data from the local schemata used by the data sources into a common data format and integrating them. The system architecture used for automating the

Table 2.1: Proteins and interactions obtained from individual data sources

Database Name	Number of proteins	Number of interactions	Number of organisms
<i>DIP</i>	16,904	43,941	104
<i>STRING</i>	270,131	473,455	110
<i>KEGG</i>	40,880	118,033	118

data integration is illustrated in figure 2.1.

The database management engine utilises transformation components in order to retrieve protein interaction data from the data sources. Each transformation component passes protein interaction data to the data management engine in the form of protein and protein interaction data structures. In the case of the DIP, which contains a relatively small amount of data, the entire content of this data source is passed to the database management engine in a single batch of proteins and protein interactions. The other transformation components, which handle data sources containing larger volumes of data, pass data to the database management engine in successive organism specific batches of data in order to optimise this process in terms of performance. The database management engine ensures the accuracy and consistency of the database when receiving data from the transformation components. At the time of writing, the eXSys database contains 487,399 proteins and 624,226 interactions among 163 organisms. A breakdown of the contributions made to the database by each of the four data sources used is displayed in table 2.1.

3. Analysis of protein interaction networks

For analysis of protein interaction networks we considered the network as a pair of sets $G = \{P, E\}$, where P is a set of nodes (proteins) and E is the set of edges (interactions) that connect two elements of P. There are different methods for identifying semantic sub-structures in a network that are essential for the network integrity. In this study we considered four features of networks; the hubs, elementary path nodes and bottleneck proteins. Details of these methods are discussed in [1]. The semantic structures identified in the protein network were ranked in the order of their

significance to categorise the most influential proteins and interactions within the network. Furthermore, we developed methods for estimating the damaging potential of each individual protein P on the ranked list by calculating the average minimum path length, the average clustering coefficient and the isolated node ratio these methods are discussed in [1].

4. e-Science toolkit

A software toolkit has been developed in order to enable e-scientists to utilise the techniques described in the previous section. This toolkit allows the retrieval of data from the eXSys database and the execution of the network analysis algorithms described above. Additionally, the results of network analyses are presented visually as illustrated in figures 4.1 and 4.2. Work is currently being performed in order to expose the two main components of the eXSys software (the eXSys database and the network analysis algorithms) as Web Services to allow the utilisation of these components within other applications.

5. Results

A significant result has been achieved by analysing the gram-positive bacteria *Bacillus subtilis* protein interaction network using data from our database. We used our algorithm methods to analyse the topological structures of the *Bacillus subtilis* network and to derive a list of the most important protein targets based on the relative damaging potential of each individual protein within the network. We found that 40% of proteins targets identified by our method were encoded by genes that are known to be essential to the survival of *Bacillus subtilis* when compared to the experimental predictions of Kobayashi and co-workers [8].

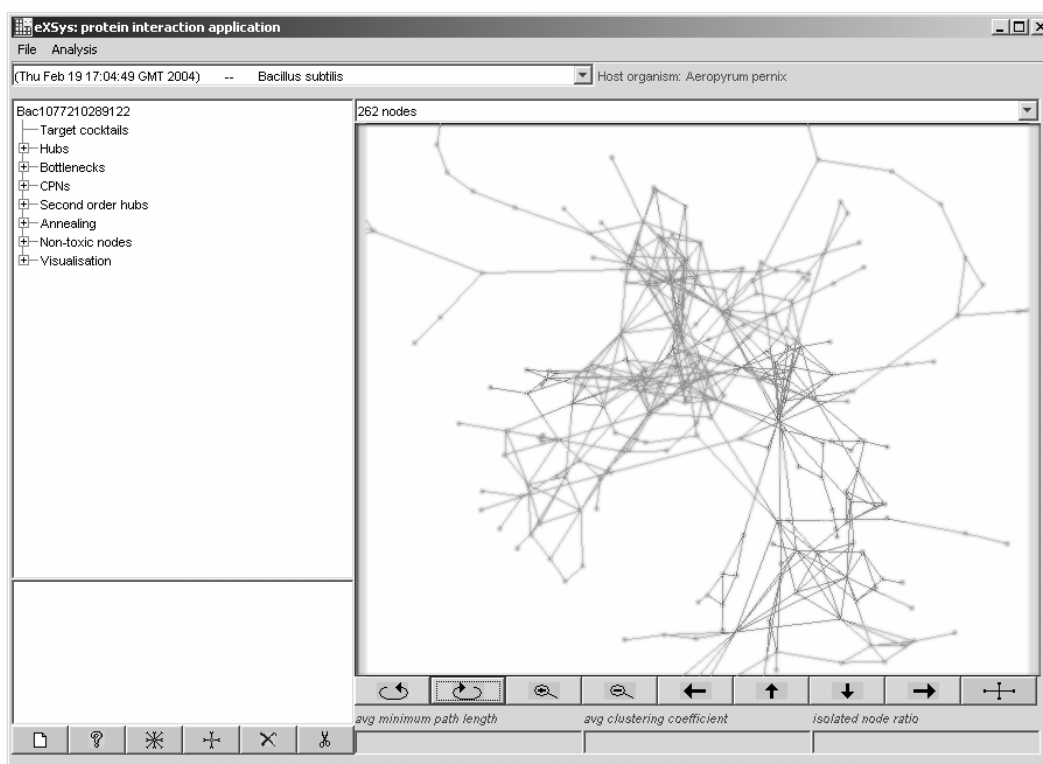


Figure 4.1: Network visualisation using the eXSys software

The eXSys software enables the visualisation of protein interaction networks as three dimensional graphs which may be manipulated visually. This allows e-scientists to visualise the characteristics of specific proteins and the results of protein deletion upon network structure. Here, part of the *Bacillus subtilis* protein interaction network is visualised.

Furthermore, we compared our protein selection methods with random protein selections and show that our selections are statistically significant [1].

6. Conclusions

The methods and software tools developed will allow e-scientists to gain significant new information about complex systems, initially in the domain of protein interaction networks, and to discover vulnerabilities within complex networks in other domains. Discovering such vulnerabilities may be of use in designing more robust networks, for example, designing more reliable software systems, designing protection policies and interventions for ecological networks.

7. References

[1] Idowu, O. C., Lynden, S. J., Brett, J., Periorellis, P., Young, M.P. and Andras, P., *Finding potential antibiotic targets by analysing protein interaction networks. Submitted to Journal of Computational Biology, 2004.*

Idowu, O. C., Lynden, S. J., Young, M.P. and Andras, P., *Bacillus Subtilis Protein Interaction Network Analysis, IEEE Computational Systems Bioinformatics Conference, Stanford, USA, California, (2004)*

<http://www.ncl.ac.uk/exsys>

[2] Barabasi, A. L. and Albert, R., 1999. *Emergence of scaling in random networks, Science, 286, 509-512.*

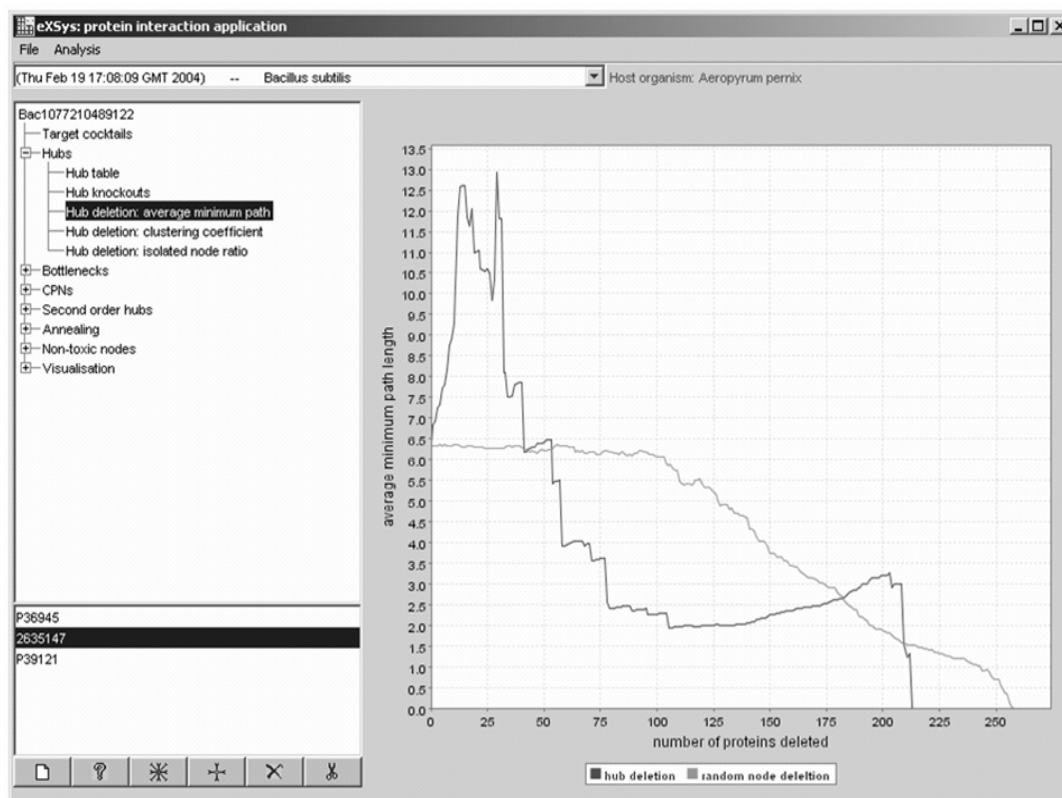


Figure 4.2: Target selection and deletion analysis using the eXSys software

The eXSys software enables the successive deletion of targets and the analysis of the effects of these deletions upon network integrity. Here, the *Bacillus subtilis* hubs are successively deleted and the effects upon the average minimum path length integrity measures are displayed. The hub deletions are compared to the effects of randomly deleted proteins, showing that the hub deletions initially have a far greater effect upon the network average minimum path length.

[4] von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., Snel, B., 2003. STRING: a database of predicted functional associations between proteins, *Nucleic Acids Res.* **31**(1), 258-261. <http://www.bork.embl-heidelberg.de/STRING>

[8] Kobayashi, k., Ehrlich, S. D., Albertini, A., Amati, G., Andersen, K. K., et al., 2003. Essential *Bacillus subtilis* genes, *Proc. Natl. Acad. Sci. USA* **100**, 4678-4683

[5] Kanehisa, M., Goto, S., Kawashima, S., and Nakaya, A. The KEGG databases at GenomeNet. 2002. *Nucleic Acids Research* **30**, 42-46

[6] Jeong, H., Mason, S.P., Barabasi A.L., and Oltvai, Z.N., 2001. Lethality and centrality in protein networks, *Nature*, **411**, 41-42

[7] Maslov, S. and Sneppen, K., 2002. Specificity and stability in topology of protein networks, *Science*, **296**, 910-913.