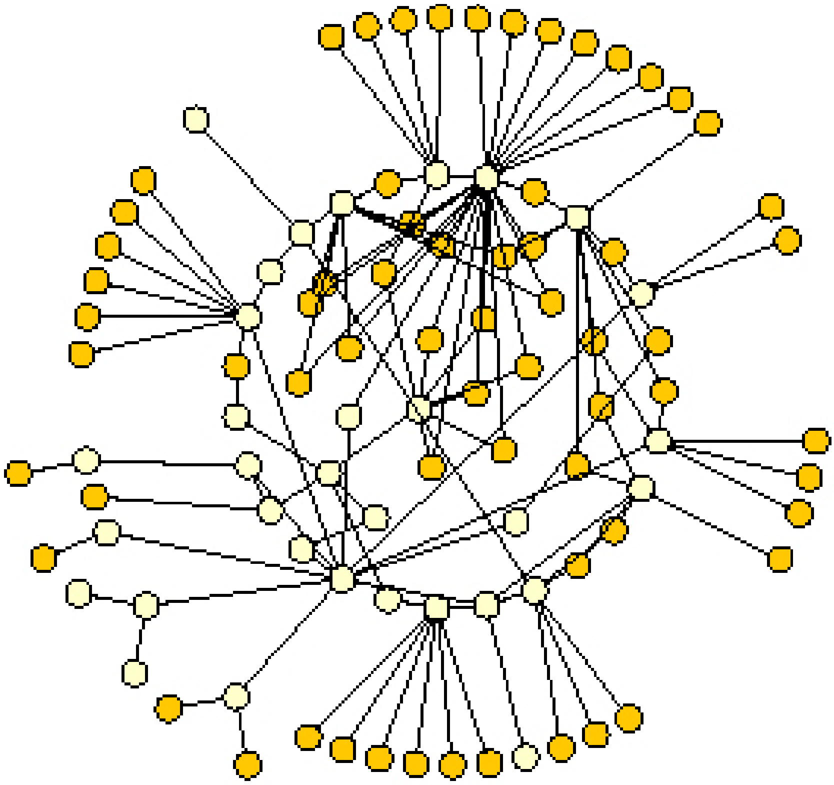
Dissertation

VTT PUBLICATIONS 732



Venkata Gopalacharyulu Peddinti

Data integration, pathway analysis and mining for systems biology

VTT PUBLICATIONS 732

Data integration, pathway analysis and mining for systems biology

Venkata Gopalacharyulu Peddinti

Department of Biomedical Engineering and Computational Science

Doctoral dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the Faculty of Information and Natural Sciences, The Aalto University School of Science and Technology, for public examination and debate in Auditorium F239a at Aalto University (Otakaari 3, Espoo, Finland) on the 14th of May, 2010 at 12 noon.

ISBN 978-951-38-7385-1 (soft back ed.)
ISSN 1235-0621 (soft back ed.)

ISBN 978-951-38-7386-8 (URL: http://www.vtt.fi/publications/index.jsp)
ISSN 1455-0849 (URL: http://www.vtt.fi/publications/index.jsp)

JULKAISIJA – UTGIVARE – PUBLISHER

VTT, Vuorimiehentie 3, PL 1000, 02044 VTT
puh. vaihde 020 722 111, faksi 020 722 4374

VTT, Bergsmansvägen 3, PB 1000, 02044 VTT
tel. växel 020 722 111, fax 020 722 4374

VTT Technical Research Centre of Finland, Vuorimiehentie 3, P.O. Box 1000, FI-02044 VTT, Finland
phone internat. +358 20 722 111, fax + 358 20 722 4374

Technical editing Maini Manninen

Edita Prima Oy, Helsinki 2010

Venkata Gopalacharyulu Peddinti. Data integration, pathway analysis and mining for systems biology. Espoo 2010. VTT Publications 732. 62 p. + app. 67 p.

Keywords systems biology, high-throughput data, data integration, data mining, visualisation, bioinformatics, conceptual spaces, network topology

Abstract

Post-genomic molecular biology embodies high-throughput experimental techniques and hence is a data-rich field. The goal of this thesis is to develop bioinformatics methods to utilise publicly available data in order to produce knowledge and to aid mining of newly generated data. As an example of knowledge or hypothesis generation, consider function prediction of biological molecules. Assignment of protein function is a non-trivial task owing to the fact that the same protein may be involved in different biological processes, depending on the state of the biological system and protein localisation. The function of a gene or a gene product may be provided as a textual description in a gene or protein annotation database. Such textual descriptions lack in providing the contextual meaning of the gene function. Therefore, we need ways to represent the meaning in a formal way. Here we apply data integration approach to provide rich representation that enables context-sensitive mining of biological data in terms of integrated networks and conceptual spaces. Context-sensitive gene function annotation follows naturally from this framework, as a particular application. Next, knowledge that is already publicly available can be used to aid mining of new experimental data. We developed an integrative bioinformatics method that utilises publicly available knowledge of protein-protein interactions, metabolic networks and transcriptional regulatory networks to analyse transcriptomics data and predict altered biological processes. We applied this method to a study of dynamic response of *Saccharomyces cerevisiae* to oxidative stress. The application of our method revealed dynamically altered biological functions in response to oxidative stress, which were validated by comprehensive *in vivo* metabolomics experiments. The results provided in this thesis indicate that integration of heterogeneous biological data facilitates advanced mining of the data. The methods can be applied for gaining insight into functions of genes, gene products and other molecules, as well as for offering functional interpretation to transcriptomics and metabolomics experiments.

Preface

This research work has been carried out at the Quantitative Biology and Bioinformatics (QBIX) Group at VTT Technical Research Center of Finland (VTT). I thank my advisor, Research Professor Matej Orešič, for his excellent guidance throughout this work and the broad range of opportunities he provided to me at work. I thank my supervisor, Professor Kimmo Kaski, Head of the Centre of Excellence, Department of Biomedical Engineering and Computational Science (BECS) of Helsinki University of Technology (TKK; called Aalto University School of Science and Technology since January 2010), for accepting me as a doctoral student and for his excellent support during this period. I thank my co-advisor, Dr. Jaakko Hollmén, Chief Research Scientist, Department of Information and Computer Science, TKK, for his cooperation and excellent advise during thesis writing.

The QBIX group has been reorganised into two teams since January 2009: Metabolomics and Biosystems Modelling; I belong to the latter. I thank the leaders of these two teams Dr. Tuulia Hyötyläinen and Dr. Marko Sysi-Aho, respectively, for their continued support to my work. I thank Technology Manager Dr. Richard Fagerström for his support. He has readily offered help and advise on a variety of issues. I am grateful to Dr. Juha Ahvenainen, former Vice President (R&D) (currently Vice President, Customer Management), Dr. Anu Kaukovirta-Norja, Vice President (R&D), and Research Professor Hans Söderlund for the excellent working facilities. I thank Mr. Atul Khanna, chairman of Tooltech Software Ltd, for his support during the early years of this work.

I thank all researchers who co-authored with me. Without their contributions to my publications this work would have been incomplete. I thank Mr. Erno Lindfors for his enthusiasm and excellent work, which have been crucial for the success of my work. Working with him has always been easy and effective. Working with Dr. Vidya Velagapudi has provided insights into biology. Her positive attitude and perseverance made working with her a motivating experience. I also thank Mr. Jussi Mattila, Dr. Jyrki Lötjönen, and other members of the Signal and image processing team at VTT Tampere for their collaboration.

The working environment at VTT has been conducive for the sort of interdisciplinary research presented in the thesis. The interdisciplinary knowledge of the members of QBIX group has created a knowledgeable workspace, and I thank each and every individual of the group. Working at QBIX group has also been full of fun. I thank Mr. Laxman Yetukuri, Dr. Catherine Bounsaythip, Mr. Han Zhao, Ms. Sandra Castillo, and Mr. Pekka Savolahti for their wonderful company. A lot of discussions with them, which included topics such as culture, life, ethics, or simply fun, at the coffee table have been revitalising. Technical discussions with my present and past team mates as well as many other researchers and faculty members at VTT and TKK, have been helpful at different instances.

I thank Professor Garry Wong and Docent Tero Aittokallio for the pre-examination of the thesis. Their valuable comments were insightful and enhanced the presentation of the thesis. I thank all my colleagues who read my thesis and provided their valuable comments. I thank Dr. Kaija Virolainen, Ms. Anna-Kaarina Hakala, and Ms. Sirpa Nygren for their help with many practicalities.

Friends at work as well as outside, including a number of Indian friends, made the social life in Finland very smooth and lively. They are, at times, unique sources of some important practical information related to living in Finland. As many of these friends are researchers in biology, bioinformatics or computer science, I also had many useful technical discussions with them.

When I was in my masters at Indian Institute of Technology (IIT) Kharagpur, my friend Dr. Murthy Chavali, then a graduate student, has introduced to me the topic of DNA computing, which subsequently led my interest to the broad variety of applications at the interface of computational sciences and biology. Support provided by Dr. G. P. Raja Sekhar at that time is also gratefully acknowledged.

I thank my wife, Subhadevi Attili, for her unconditional love and trust. I thank my parents and all other family members for their love, affection, and support. What I am today is the net result of my actions in the past and how they were supported by family, friends, teachers, and possibly even people I do not know. Although the names of all of those who played important roles in my life are not mentioned, they are already part of what I am today, and I am indebted to their support.

List of publications

This thesis consists of introductory part and the following publications

Article I. Peddinti V. Gopalacharyulu, Erno Lindfors, Catherine Bounsaythip, Teemu Kivioja, Laxman Yetukuri, Jaakko Hollmén, and Matej Orešič. Data integration and visualization system for enabling conceptual biology. *Bioinformatics*, 21 Suppl 1:i177–i185, Jun 2005.

Article II. Peddinti V. Gopalacharyulu, Erno Lindfors, Jarkko Miettinen, Catherine Bounsaythip, and Matej Orešič. An integrative approach for biological data mining and visualisation. *Int. J. Data mining and Bioinformatics*, 2(1):54–77, Jan 2008.

Article III. Peddinti V. Gopalacharyulu, Erno Lindfors, Catherine Bounsaythip, and Matej Orešič. Context dependent visualization of protein function. In Juho Rousu, Samuel Kaski, and Esko Ukkonen, editors, *Probabilistic Modeling and Machine Learning in Structural and Systems Biology*, pages 26–31, Tuusula, Finland, Jun 2006.

Article IV. Catherine Bounsaythip, Erno Lindfors, Peddinti V. Gopalacharyulu, Jaakko Hollmén, and Matej Orešič. Network-based representation of biological data for enabling context-based mining. In Catherine Bounsaythip, Jaakko Hollmén, Samuel Kaski, and Matej Orešič, editors, *Proceedings of KRBIO'05, International Symposium on Knowledge Representation in Bioinformatics*, pages 1–6, Espoo, Finland, Jun 2005. Helsinki University of Technology, Laboratory of Computer and Information Science.

Article V. Peddinti V. Gopalacharyulu, Vidya R. Velagapudi, Erno Lindfors, Eran Halperin, and Matej Orešič. Dynamic network topology changes in functional modules predict responses to oxidative stress in yeast. *Mol. BioSyst.*, 5:276–287, 2009.

Roman numerals are used to refer to these articles.

Author's contributions

The research presented in this thesis consists of two themes: data integration and context-sensitive visualisation (**Articles I–IV**), and dynamic topology of integrated networks (**Article V**). Author's contributions to each of these papers have been summarised below. Contributions of other authors are mentioned when necessary.

Article I presents a heterogeneous data integration system for the integration and visualization of multiple types of biological interaction data and demonstrates its usage with two applications: multiple pathway retrieval and protein neighbourhood search. The author designed the system, performed data modelling, developed the schemas, obtained relevant data, developed parsers, implemented the database back-end and developed the semantics for the integration of heterogeneous biological entities, and provided guidance to the developement of the overall system. Author wrote the first draft of the paper which was then improved with the contributions of other authors. Erno Lindfors (EL) implemented the integration of biological entities, representation of integration results using complex networks, the Sammon's mapping algorithm, Graphical User Interface client software and wrote these parts of the manuscript.

Article II presents the data integration system with addition of new methods. It presents three applications: system-wide metabolic network and the study of its topological properties, exploration of properties and relationships of a specific set of proteins, and combined visualization and exploration of a Type 1 Diabetes gene expression data in mouse together with related pathways and ontologies. Author developed the ideas concerning integration of gene expression data to networks and performed the analyses and wrote the manuscript. EL developed the ideas for topology study and performed analyses and implemented middle-tier and user interface and wrote some parts of the manuscript. The author and EL equally contributed to this work. Jarkko Miettinen (JM) implemented projection algorithms and improved the client and middle-tier software design and code.

Article III presents the application of the data integration framework for context-dependent visualization of protein function using network representation and nonlinear projection methods based on Curvilinear Distance Analysis. The author performed the analyses and wrote the methods and results parts of the paper which was then improved by all authors under the coordination of Matej Orešič (MO).

Article IV presents the details of the networks and the distance metrics. It demonstrates with experiments how judicious use of various distance functions can allow emergence of context. The author mainly provided biological details and data, and contributed to the manuscript which was mainly written by Catherine Bounsaythip (CB) and EL.

Article V presents an integrative bioinformatics method called Topological Enrichment Analysis of Functional Subnetworks (TEAFS). Author developed the main ideas of the method, implemented some parts of the method, performed the analyses and wrote the manuscript. Vidya R. Velagapudi (VRV) performed *in vivo* metabolic profiling experiments and data analysis, and wrote the experimental methods and biological details in the manuscript. Author and VRV equally contributed to this work. EL performed biological network construction, provided help with topological analysis, implemented the statistical test and contributed to the writing. Eran Halperin (EH) provided ideas for the statistical test. MO initiated and coordinated the project and contributed to the writing of the paper.

Goals of the thesis

The goals of this thesis are

- to develop integrative bioinformatics tools and methods to integrate heterogeneous biological data
- to produce knowledge, *e.g.,* testable hypotheses, from these tools in an exploratory manner
- to apply these tools and methods in mining newly generated experimental data

Contributions presented in this thesis

Following contributions have been made through this thesis.

- A technical framework for storage and retrieval of biological data achieved through XML and relational databases.
- The technique of data traversals, achieved by curation of *maps* database, as the basis for heterogeneous data integration.
- A platform called *MegNet* for enabling context-sensitive mining of heterogeneous biological data through the usage of conceptual spaces as the framework for knowledge representation.
- A new method called Topological Enrichment Analysis of Functional Subnetworks (TEAFS) for studying the dynamic activity of biological process modules, and its application to studying the dynamic response of *Saccharomyces cerevisiae* to Oxidative stress.

Contents

**Article V is not included in the PDF version.
Please order the printed version to get the complete publication
(http://www.vtt.fi/publications/index.jsp).**

List of Abbreviations

Abbreviation	Expansion
A	absent
API	Application Programming Interface
BioGrid	General Repository of Biological Interaction Datasets
BIND	Biomolecular Interaction Database
CCA	Curvilinear Component Analysis
CDA	Curvilinear Distance Analysis
cDNA	Complementary DNA
DBMS	Database Management System
DIP	Database of Interacting Proteins
DNA	Deoxyribonucleic Acid
DOM	Document Object Model
DTD	Document Type Definition
EC	Enzyme Commission
EMBL	European Molecular Biology Laboratory, EMBL Nucleotide sequence database
FDR	False Discovery Rate
GC	Gas Chromatography
GDS	GEO Data Set
GEO	Gene Expression Omnibus
GI	GenInfo sequence identifier
GNEA	Gene Network Enrichment Analysis
GO	Gene Ontology
GSEA	Gene Set Enrichment Analysis
GPL	GEO Platform
GSM	GEO Sample
GSE	GEO Series
HPLC	High Performance Liquid Chromatography
JDBC	Java Database Connectivity
KEGG	Kyoto Encyclopedia of Genes and Genomes
KGML	KEGG Markup Language
M	Marginal
MINT	Molecular Interaction Database
MM	Mismatch
mRNA	messenger RNA
MS	Mass Spectrometry, Mass Spectrometer

Abbreviation	**Expansion**
NOD	Non Obese Diabetic
P	Present
PDB	Protein Data Bank
PIR	Protein Information Resource, PIR-international protein sequence database
PM	Perfect Match
PSI	Protein Standards Initiative
PSI-MI	PSI-Molecular Interaction format
RDF	Resource Description Framework
RNA	Ribonucleic Acid
SAGE	Serial Analysis of Gene Expression
SAX	Simple API for XML
SBML	Systems Biology Markup Language
SOAP	Simple Object Access Protocol
SOFT	Simple Omnibus Format in Text
SQL	Structured Query Language
TCA cycle	The citric acid cycle or The tricarboxylic acid cycle
TEAFS	Topological Enrichment Analysis of Functional Subnetworks
TRANSFAC	Database of Transcription Factors
TRANSPATH	Database of Signal Transduction Pathways
TrEMBL	Translated EMBL
T1D	Type 1 Diabetes
UMLS	Unified Medical Language System
UniProt	Universal Protein Resource
UPLC	Ultra Performance Liquid Chromatography
XML	eXtensible Markup Language
XPath	XML Path Language
XQuery	XML Query Language
XSD	XML Schema Definition
Y2H	Yeast Two-Hybrid method

Chapter 1

Introduction

Systems biology aims to study biological systems at system level (Kitano, 2002). Integrative bioinformatics provides tools for systems biology. The necessary goal of integrative bioinformatics is to provide platforms and methods for carrying out systems biology analysis. Ge et al. (2003) noted that, "high-throughput data integration is needed in systems biology approaches". This thesis contributes methods for data integration, visualisation, mining, and study of network dynamics, which are important elements of systems biology Figure 1.1.

Post genomic biology is characterized by rapid accumulation of information and thus biological research can be viewed as information science (Schena et al., 1998). Information is increasing more swiftly than humans resources such as time and cognitive processing power. This forces humans to narrow their focus in processing the information, which in turn causes diminishing awareness. As an example, the literature documenting Raynaud's disease and fish oils were mutually isolated in the sense that the authors and readers of one literature were not aware of the other. Much later, a cure for Raynaud's disease with dietary fish oils was found, using a literature-based approach covering both areas (Swanson, 1986). This demonstrates the power of data integration. Thus the methods for efficient retrieval and presentation of results can benefit researchers.

Cytoscape (Shannon et al., 2003; Killcoyne et al., 2009) is a general network visualization, data integration, and analysis software, which has been mainly developed with the modelling requirements of systems biology in mind. The core of Cytoscape mainly features powerful layout algorithms for visualising networks and is quickly becoming a *de facto* standard for the visualisation of biological networks, while its flexible plug-in architecture brings the real power via community-based development of useful plug-ins. However, Cytoscape does not offer data management capabilities. Biological data management and integration has also attracted significant amount of research (Lacroix and Critchlow, 2003). There are three major approaches in this area: data warehousing approach, distributed or federated approach, and mediator approach. The data warehousing approach involves assembling data sources into a centralised system with a global data schema and an indexing system for integration and navigation. In the federation approach, underlying data sources remain autonomous, and the federated system maintains a common data model and makes use of schema mapping to translate heterogeneous source database schema to the target schema for integration. The mediator approach introduces a mediator layer, a collection of software components performing integration, to decouple the underlying heterogeneous distributed data sources and the client layer. Many popular and important biological data integration systems are discussed in Lacroix and Critchlow (2003). But these systems are limited to providing web based access to multiple reference databases.

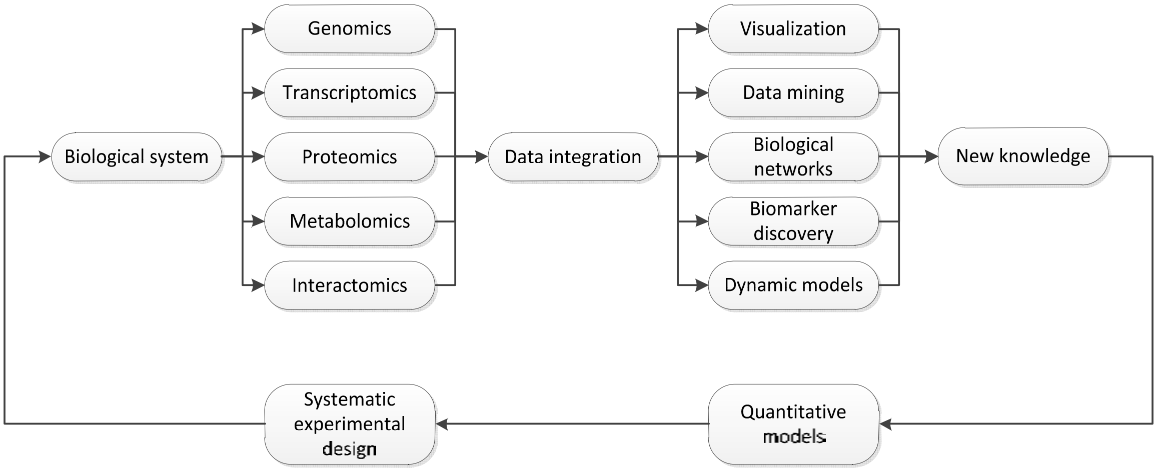


Figure 1.1: Systems biology can be viewed as a combination of omics technologies, data integration, analysis, mining, and modelling, often involving use of these techniques iteratively over hypothesis driven systematic experimental design to gain increased understanding of the structure and dynamics of the biological systems. High throughput omics technologies provide the measurements for systems biology. Integrative bioinformatics starts with the integration of multiple data sets from one or more omics and also possibly from multiple organisms, and forms the basis for systems biology analysis. Systems biology analyses include data mining, visualisation, biological networks, and dynamic modelling. The new knowledge generated by these analyses would enable us to build quantitative models. The hypothesis generated by the analysis of these models drive the design of more experiments to gain increased understanding of the biological systems.

The availability of high-throughput data collection techniques of modern biology introduce some new problems. First, there are many false positive findings and reproducibility is poor *i.e.,* a biological sample analysed by using a single experimental technique at different times or laboratories often lead to unidentical results (Ge et al., 2003; Ein-Dor et al., 2005; Irizarry et al., 2005; Tan et al., 2003). Meta-analyses of multiple data sets or evidences from multiple types of biological experiments may improve the statistical power of the analysis. Second, various types of biological activities or interactions within an organism and between an organism and its environment do not happen in isolation. Biological function is a net result of simultaneous activities and interactions of various types (Kanehisa and Bork, 2003; Ideker et al., 2001a,b; Ge et al., 2003; Papin and Palsson, 2004). Thus the integrated modelling of the biological systems is very important.

This thesis addresses these problems. First, it presents bioinformatics methods to visualise biological interactions of different types in an integrated manner. Next, it presents methods to facilitate advanced context-sensitive mining of the integrated network data. Finally, it presents mining of dynamic topological changes of functional modules in integrated networks in response to specific interventions. The methods presented in this thesis can be applied for prediction of gene, gene product or metabolite functions, and to associate experimental phenotypes with genotypes *i.e.,* for interpretation of transcriptomics and metabolomics data in terms of molecular level patterns or changes.

Organisation of the thesis

The research presented in this thesis consists of integrative bioinformatics methods under two themes. The first theme is that of a bioinformatics software platform which embodies methods for integration of heterogeneous biological data: a variety of interaction, annotation, and molecular measurement data (chapter 2). The methodological details of this software platform are presented in section 3.1. The platform achieves integration across different biological data types using data traversals (section 3.1.4). The resulting information forms the basis for context-sensitive data mining (section 3.1.6), which draws on the concepts from the theory of conceptual spaces. Some results of exploratory data mining using this platform are presented in section 4.1. The second theme is the study of dynamic changes of functional modules in an integrated network (section 3.2). An application of this strategy with a study of dynamic topological response of oxidative stress in *Saccharomyces cerevisiae* is presented in section 4.2. Summary and concluding remarks follow in chapter 5.

Chapter 2

Biological data

Cells consist of different types of biomolecule. Reductionist approach of molecular biology deals with studying the properties and roles of individual molecules and produced large amounts of useful knowledge. A discrete biological function, however, can not be attributed to an individual molecule, but to a complex web of interactions between a set of molecules. Therefore, describing biological systems requires explaining how they arise from interactions among components in the cell (Hartwell et al., 1999; Ge et al., 2003). Availability of complete genome sequences of several organisms (Goffeau et al., 1996; Blattner et al., 1997; Adams et al., 2000; Lander et al., 2001; Venter et al., 2001; Waterston et al., 2002) has opened doors for development of high-throughput *omics* technologies.

Figure 2.1 provides an overview of some *omics* experimental techniques for measuring different biological data types, which are elaborated in this chapter. Gene expression microarrays (Schena et al., 1995; Lockhart et al., 1996) and Serial Analysis of Gene Expression (SAGE) (Velculescu et al., 1995), which enable us to measure the abundances of thousands of gene transcripts simultaneously, have empowered the first *omics* discipline known as transcriptomics. The development of proteomics, large scale study of proteins, followed (Patterson and Aebersold, 2003). Mass spectrometry based protein purification (Rigaut et al., 1999; Aebersold and Mann, 2003) and yeast two hybrid analysis (Ito et al., 2001) have enabled investigating protein-protein interactions in high-throughput manner (Shoemaker and Panchenko, 2007a). The *omics* technologies have contributed to rapid accumulation of knowledge such as gene and gene product annotations, biomolecular interactions. Simultaneously, the availability of genome sequences also enabled development of computational algorithms for sequence analyses (Altschul et al., 1990; Thompson et al., 1994; Durbin et al., 1998), which also helped to rapidly annotate new sequence data, and predict the structure and interactions (Shoemaker and Panchenko, 2007b; Marcotte et al., 1999). Finally, text mining (Skusa et al., 2005) also helped to retrieve important molecular interaction information, and careful manual literature curation led to higher quality information as compared to high-throughput data (Reguly et al., 2006; Matys et al., 2006). In the spirit of genomic data sharing, many types of the biological data have been made accessible through world wide web.

These web-accessible databases and their associated search and mining tools are primary resources serving thousands of biology researchers worldwide. These tools allow researchers to effectively mine the databases and answer one's biological questions (Kanehisa and Bork, 2003). The databases cover a wide range of information including literature, sequences and annotations of genes and gene products, and a variety of molecular interactions such as biochemical reactions, transcriptional regulatory interactions, signal transduction pathways, to name a few. This chapter introduces some of the commonly used databases for miroarray gene expression profiles, protein-protein

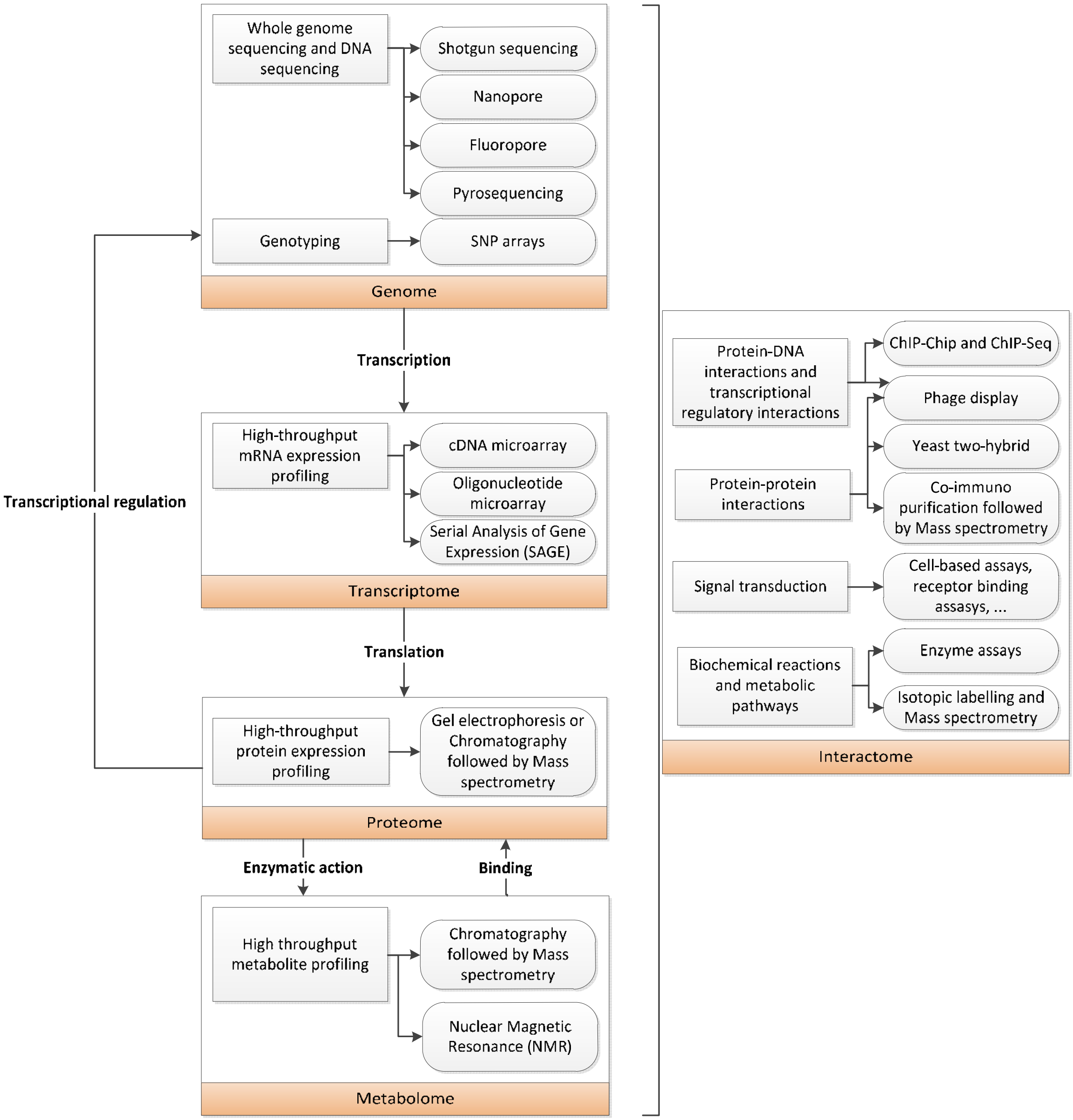

Whole genome sequencing and DNA sequencing
Shotgun sequencing
Nanopore
Fluoropore
Pyrosequencing
Genotyping
SNP arrays
Genome
Transcription
High-throughput mRNA expression profiling
cDNA microarray
Oligonucleotide microarray
Serial Analysis of Gene Expression (SAGE)
Transcriptome
Translation
High-throughput protein expression profiling
Gel electrophoresis or Chromatography followed by Mass spectrometry
Proteome
Enzymatic action
Binding
High throughput metabolite profiling
Chromatography followed by Mass spectrometry
Nuclear Magnetic Resonance (NMR)
Metabolome
Transcriptional regulation
Protein-DNA interactions and transcriptional regulatory interactions
ChIP-Chip and ChIP-Seq
Phage display
Yeast two-hybrid
Protein-protein interactions
Co-immuno purification followed by Mass spectrometry
Signal transduction
Cell-based assays, receptor binding assasys, ...
Biochemical reactions and metabolic pathways
Enzyme assays
Isotopic labelling and Mass spectrometry
Interactome

Figure 2.1: Various types of data in molecular biology and the experimental techniques employed for obtaining the data (by no means exhaustive). High-throughput techniques capable of measuring all or at least a large number of components (several hundreds to thousands) simultaneously are known as *omics* techniques. Generally the word *interactome* is used to refer to the collection of protein-protein interactions, but in this thesis we also include other types such as metabolic networks and transcriptional regulatory interactions into the *interactome* category for convenience, as we repeatedly refer to all these types of interactions.

interactions, metabolic interactions, transcriptional regulatory interactions, and signal transduction networks.

GenBank (Benson et al., 2000, 2010) is a sequence database that stores all known DNA sequences gathered by direct submission of sequence data from individual laboratories and from large-scale sequencing projects. There are two other major DNA sequence databases namely EMBL nucleotide sequence database (Stoesser et al., 1999; Kulikova et al., 2007), and DDBJ (Tateno et al., 2002; Kaminuma et al., 2010), and the data among these three databases are synchronised. The Ensembl project (Hubbard et al., 2007; Flicek et al., 2010) offers an integrated source of genome sequences and annotations for a comprehensive set of chordate genomes with a particular focus on human, mouse, rat, zebrafish etc. The University of California Santa Cruz (UCSC) Genome Browser Database (Karolchik et al., 2003; Rhead et al., 2010) is a source for genome sequence and annotation data. UCSC Genome Browser, is a tool associated with the UCSC database that provides rapid visualization and querying of the data. The annotations provided by these genome databases include include mRNA and expressed sequence tag (EST) alignments, gene predictions, cross-species homologies, highlevel maps, single nucleotide polymorphisms (SNPs) and so on. Besides these general genomic databases, there are organism specific genomic databases for model organisms. The Saccharomyces Genome Database (SGD) (Cherry et al., 1998; Engel et al., 2010) is a database for the molecular biology and genetics of the yeast *Saccharomyces cerevisiae* that provides functional annotations, mapping and sequence information, protein domains and structure, expression data, mutant phenotypes, physical and genetic interactions and the primary literature from which these data are derived. FlyBase (Gelbart et al., 1997; Drysdale et al., 2008) is a database of genetic and genomic data concerning fruit flies of which *Drosophila melanogaster* is an extensively studied model organism. FlyBase is populated with information from a variety of sources ranging from large-scale genome projects to the primary research literature. FlyBase provides access to information on gene models, molecular classification of gene product functions, mutant phenotypes, mutant lesions and chromosome aberrations, gene expression patterns, transgene insertions, and anatomical images. WormBase (Stein et al., 2001; Harris et al., 2010) is a central data repository for nematodes of which *Caenorhabditis elegans* is an extensively studied model organism. WormBase includes genomic sequences, gene predictions and orthology assignments from a range of related nematodes and relies on manual curation of information from the corpus of *C. elegans* literature.

Transcription factors are proteins that are vital for the transcriptional regulation of gene expression. A transcription factor has a DNA binding domain which can bind to a particular region in the DNA sequence of a gene, called the binding site, and helps in enhancing or inhibiting the expression of the gene (Latchman, 1997). TRANSFAC database (Wingender et al., 2000; Matys et al., 2006) primarily provides information about entities involved in the transcriptional regulation such as transcription factors, binding sites and genes among a variety of other related information.

Living cells interact with their environment by exchanging a variety of signals. Signaling pathways of the receiver cells forward the signals to the nucleus through cascades of interactions and trigger the appropriate adaptation of the genetic program. The TRANSPATH database (Schacherer et al., 2001; Krull et al., 2006) provides information about signal transduction pathways involved in the transcriptional regulation of gene expression via regulating the activity of the transcription factors.

The Universal Protein Resource (UniProt) (Bairoch et al., 2005; Apweiler et al., 2010) provides information about protein sequences and functional information. The central database in Uniprot, termed UniProt Knowledgebase, provides accurate, consistent and rich sequence and functional annotations and consists of two sections: UniProt/Swiss-Prot and UniProt/TrEMBL. UniProt/Swiss-Prot consists of manually curated protein functional information, resulting from literature information extraction and curator-evaluated computational analysis. UniProt/TrEMBL consists of protein sequences translated from EMBL gene sequences (Stoesser et al., 1999; Kulikova et al., 2007) and annotated with computational annotation tools, pending manual curation.

Database of Interacting Proteins (DIP) (Xenarios et al., 2002) is a database of manually curated protein-protein interactions. A curator enters each interaction entry into the database after manually reading the publication reporting an experimentally verified interaction. This is intended to be a comprehensive and integrated tool for browsing and efficiently extracting information about protein interactions and interaction networks in biological processes. DIP provides access to combined information from multiple observations and experimental techniques, from multiple organisms, as well as to networks of interacting proteins. Each interaction entry in the DIP database contains information about the protein domains and range of amino acids involved in the interaction, and the corresponding experiments. The interactors are identified by Swissprot (Apweiler et al., 2010), PIR (Barker et al., 1998), or GenBank (Benson et al., 2000) accession numbers and each interactor entry contains information about the organism, function, superfamily, cellular location and so on.

The Biomolecular Interaction Database (BIND) (Bader et al., 2003) stores pairwise interactions between biological 'objects' which could be protein, RNA, DNA, molecular complex, small molecule, photon (light) or gene. Moreover, it contains higher level functional structures called molecular complexes and pathways which are collections of the pairwise interactions with some additional data. The minimum amount of information required to define an interaction is a description of the interacting objects and a publication reference to PubMed (Wheeler et al., 2007). Data in BIND is primarily obtained via submissions of individual contributors across the world. However, it also incorporates interaction data imported from other databases such as PDB, and a number of large-scale cell mapping studies using yeast two hybrid, mass spectrometry, genetic interactions and phase display. SeqHound is a data integration system (Michalickova et al., 2002) that provides extensive C, C++, and Perl application programming interfaces (API) for data in BIND. SeqHound system provides also functions to link the biological objects with other biological databases in public domain.

The Molecular Interaction database (MINT) (Zanzoni et al., 2002; Ceol et al., 2010) stores information about experimentally verified molecular interactions extracted from publications from peer-reviewed journals. The main focus is on physical interactions between proteins. Genetic or computationally inferred interactions are not included in MINT. MINT includes an additional database called HomoMINT (Persico et al., 2005), which is a database of interactions between human proteins inferred from interactions between orthologous proteins in model organisms. A large number of MINT data comes from large scale, genome wide experiments, although curating data from low-throughput published experiments is given emphasis. Each interaction entry contains reference to Swiss-Prot/TrEMBL protein accession number (Apweiler et al., 2010) for the interactor and contains the experimental information and pubmed reference (Wheeler et al., 2007) for describing the experimental conditions and other properties of the interaction.

BioGrid (Stark et al., 2006; Breitkreutz et al., 2008) is a database of protein and genetic interactions. It is aimed to be a generic repository providing comprehensive information on molecular interactions in several organisms such as *Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Homo sapiens*. It currently hosts protein-protein interaction data from high-throughput experiments such as yeast two-hybrid (Y2H) method (Ito et al., 2001) and mass spectrometry analysis of purified protein complexes (Rigaut et al., 1999). Additionally, the BioGrid team also compiles interaction data by extensive manual curation of literature. Literature curated data for *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* have already been added to BioGrid and curation efforts for other organisms are underway (Reguly et al., 2006; Breitkreutz et al., 2008).

The Kyoto Encyclopedia of Genes and Genomes (KEGG) system (Kanehisa et al., 2004) consists of three main components: the genomic space (KEGG GENES), the chemical space (KEGG LIGAND), and network space (KEGG PATHWAY). The KEGG GENES is a collection of gene catalogues for completely or partially sequenced genomes, compiled by automatically extracting information from databases such as NCBI GenBank, and RefSeq (Wheeler et al., 2007). The KEGG LIGAND is further divided into several components including COMPOUND, GLYCAN, REACTION, ENZYME, and so on (Goto et al., 2002). The COMPOUND database contains manually entered and computationally verified chemical structures of known metabolic compounds, and some pharmaceutical and environmental compounds. The GLYCAN database consists of carbohydrate structures, a few hundreds of which were manually entered and the rest derived from CarbBank project (Doubet et al., 1989). The REACTION database contains reaction formulae for enzymatic reactions, the reactants of which are represented in COMPOUND or GLYCAN databases. The ENZYME database contains enzyme nomenclature. Each enzyme is identified by an Enzyme Commission (EC) number (Webb, 1992; Tipton and Boyce, 2000), which can be linked to other public databases such as UniProt (Bairoch et al., 2005). The KEGG PATHWAY database is a collection of manually drawn diagrams, called KEGG reference pathway diagrams (maps), each of which corresponds to a known network of functional significance. Moreover, PATHWAY database also contains organism-specific pathways, which are automatically generated by superimposing genes in given organisms. The KEGG pathways are provided in an XML based markup language called KGML. Each metabolic reaction in a KEGG metabolic pathway is linked with one entry in the REACTION database, and the enzymes in the enzymatic reactions can be linked to the databases in the genome space as well as to other public databases via EC numbers.

iND750 is a manually reconstructed genome-scale metabolic model describing *Saccharomyces cerevisiae* metabolism (Duarte et al., 2004) with 750 genes, their transcripts, proteins and reactions. Manual reconstruction process involves curating reaction lists based on information from genome annotations, biochemical pathway databases, biochemistry textbooks, and publications (Förster et al., 2003). All reactions in iND750 model are elementally and charge balanced, and compartmentalised to eight cellular locations: extracellular space, cytosol, mitochondrion, peroxisome, nucleus, endoplasmic reticulum, Golgi apparatus, and vacuole. Similar semi-automated manual curation has been employed to construct the first consensus metabolic network for yeast (Herrgård et al., 2008), global human metabolic network (Duarte et al., 2007), and so on.

Gene Expression Omnibus (GEO) (Edgar et al., 2002; Barrett et al., 2005) stores a variety of high-throughput molecular abundance data of which microarray gene expression data is a major data type. The data in GEO is organized into GEO Platforms (GPL), GEO Samples (GSM), GEO Series (GSE) and GEO Data sets (GDS). A Platform describes the set of elements that can be detected and quantified in the experiment. A Sample describes a single hybridization or experimental condition. A Series is a group of related Samples that make up one single study. A Data set is an assembly of biologically meaningful Samples that are statistically comparable. Of these, GPL, GSM and GSE are direct submissions of contributors, while GDS is a curated collection.

Gene Ontology (GO) (Ashburner et al., 2000; Berardini et al., 2010) consists of three independent ontologies: Biological process, Molecular function, and Cellular component to describe the roles of genes and proteins in eucaryotes. GO is built on the premise that a large fraction of the genes specifying core biological functions are shared by all eucaryotes. It is aimed to be a dynamic controlled vocabulary applicable to all eucaryotes even as our knowledge of gene and protein roles in cells continuously evolves.

Although the high-throughput experimental techniques of modern molecular biology empower us to measure multiple components of a biological system simultaneously, they often produce data that is inferior in quality to low-throughput techniques. On the one hand, high-throughput techniques may produce large number of false positives, meaning that not all findings are necessarily correct. On the other hand, they may also exhibit large number of false negatives or detection biases, meaning that they may miss some true phenomena, leading to the lack of coverage or comprehensiveness in the findings. von Mering et al. (2002) performed a detailed comparison of multiple high-throughput techniques as well as a computational approach commonly employed for the study of protein-protein interactions. They estimated that more than half of all high-throughput protein-protein interaction data are false positives. While the estimated lower-bound to the number of protein-protein interactions is 30,000, the number of interactions supported by more than one method is only approximately 2,400, which demonstrates the sparing coverage of the methods employed (von Mering et al., 2002). For these reasons, manual curation of published literature has, despite the practical difficulties, received significant interest. In order to compile high-quality data, manual literature curation is expected to consider high-confidence data coming from low-throughput techniques, and perhaps also additionally supported by multiple independent studies. However, a recent study that has systematically compared the quality of a literature curated data has speculated that the quality of the literature curated yeast protein interaction data from BioGrid (Reguly et al., 2006) is at the most as good as or even inferior to high-throughput yeast two-hybrid (Y2H) data (Cusick et al., 2009). On the other hand, high-throughput techniques such as Y2H method are also improving, and it is increasingly believed that the little overlap among different high-throughput protein interaction studies is due not to the false positives, but to the false negatives (Lemmens et al., 2010). However, still the coverage of such data is quite little. For instance, while reporting a newly produced high-quality comprehensive binary Y2H protein interaction map for *Saccharomyces cerevisiae*, Yu et al. (2008) indicated that three proteome-level Y2H studies (*i.e.,* Uetz et al. (2000); Ito et al. (2001); Yu et al. (2008)) taken together, only account for approximately 20% of the empirically estimated protein binary interactions in *Saccharomyces cerevisiae*. Y2H interaction maps have also been generated for other model organisms and humans (see references cited by Yu et al. (2008)), and their quality and coverage are similar to those of *S. cerevisiae* interaction maps. Global metabolic network reconstruction approaches described earlier (Duarte et al., 2004; Herrgård et al., 2008; Duarte et al., 2007) presumably produce high-quality metabolic networks but they are not complete and continuously keep growing. Comprehensive system-level data of high-quality is a key ingredient of systems biology (Kitano, 2002). Similarly, efficient computational tools to effectively handle current and future high-throughput data and turn them into knowledge are equally important.

Most of the databases described in this chapter provide easy access to web interfaces and tools for mining the data. However, these tools as well as the underlying data formats are different from each other and thus very diverse. A platform for accessing all these databases in a unified fashion as well as for performing advanced data mining of the resulting data is presented in the next chapter.

Chapter 3

Methods

3.1 *MegNet* platform for biological data integration

Although high-throughput experimental techniques of molecular biology and biochemistry produce increasing amounts of data such as gene, protein, and metabolite expression, as well as interactions between them, characterising a sample's phenotype in the context of environment or experimental condition remains a challenge. The aim of data integration is to address this challenge by providing links between a variety of these data.

By constructing a system to represent experimental phenotypes and environmental context of samples in GEO (Barrett et al., 2005) with annotations from Unified Medical Language System (UMLS) (Bodenreider, 2004), and mining the data across multiple datasets representing similar biological contexts, Butte and Kohane (2006) showed that a large set of phenome-genome and envirome-genome relations could be retrieved within a public repository of transcriptome measurements (GEO). Orešič et al. (2004) explored integrated analysis of gene, protein and metabolite expression profiles, attempted to interpret the results in the biological context using pathways, and emphasised the need for data traversals.

We have developed a bioinformatics platform, a consolidation of multiple heterogeneous molecular biology databases, and a visualisation software called *MegNet* for automatic integrative mining of these data (**Article I**, **Article II**). This section explains the technical details of the platform and *MegNet*. Henceforth, we use the words *MegNet* and *data integration platform* interchangeably to represent the database system and the visualisation software together.

Fundamentally *MegNet* system has been developed to achieve integration of heterogeneous biological data by enabling traversals across different data sources. It enables traversals across protein-protein interactions, transcriptional regulation reactions, metabolic pathways, metabolic models, signal transduction pathways, biological ontologies, and molecular profile data such as gene expression measurements (Figure 3.1). *MegNet* represents the integrated data as networks. Furthermore, *MegNet* enables context-based visualisation of the integrated networks by building a conceptual space representation (Gärdenfors, 2000) and making use of the dimensionality reduction techniques (Carreira-Perpiñan, 1997) to visualize the similarity structure in a low-dimensional space, typically a two-dimensional plot.

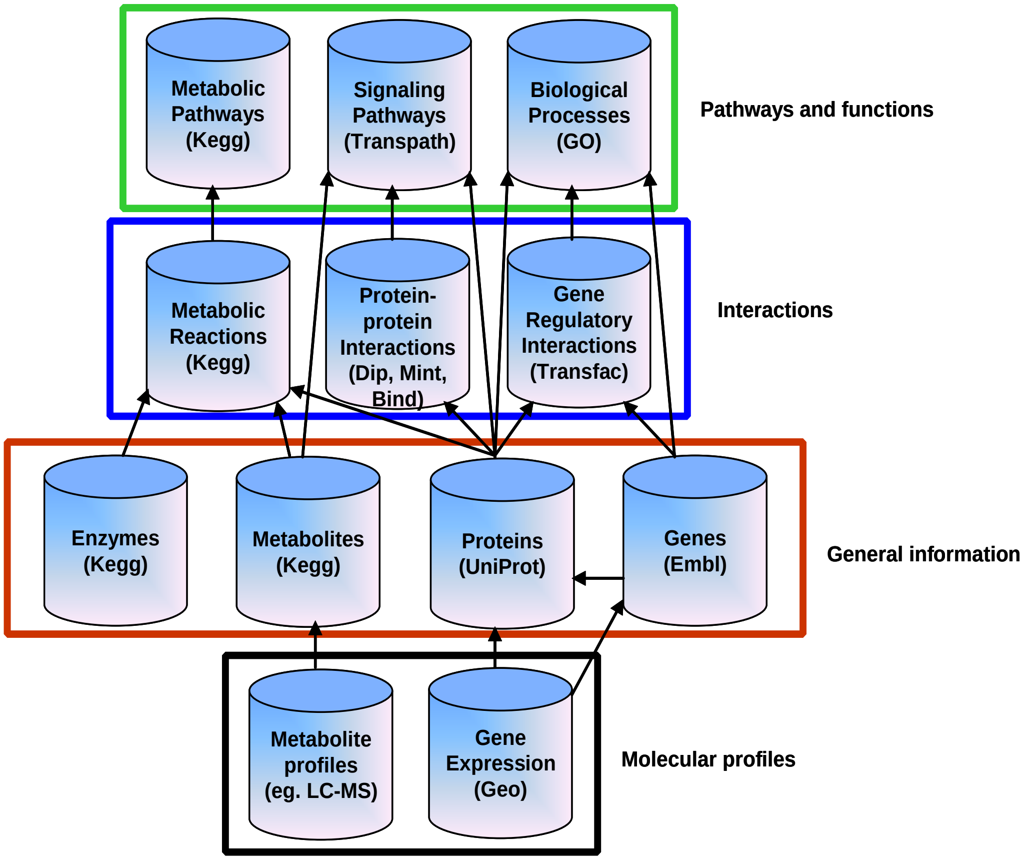

Metabolic Pathways (Kegg)
Signaling Pathways (Transpath)
Biological Processes (GO)
Pathways and functions
Metabolic Reactions (Kegg)
Protein-protein Interactions (Dip, Mint, Bind)
Gene Regulatory Interactions (Transfac)
Interactions
Enzymes (Kegg)
Metabolites (Kegg)
Proteins (UniProt)
Genes (Embl)
General information
Metabolite profiles (eg. LC-MS)
Gene Expression (Geo)
Molecular profiles

Figure 3.1: Multiple molecular biology databases provide descriptions of biological systems at different levels of abstraction. Some common biological information, along with names of primary databases providing information in that domain are provided. Four levels of biological information are indicated by boundaries: molecular profiles (black), general information about molecules (red), interactions (blue), and biological pathways and functions (green).

3.1.1 Databases

A database is merely a collection of information that exists over a long period of time. A Database Management System, DBMS, is a software system that supports storage of large databases, provides efficient access to the data through powerful query languages, supports atomic and independent execution of concurrent transactions, and supports durability—the ability to recover from failures or errors (Garcia-Molina et al., 2002).

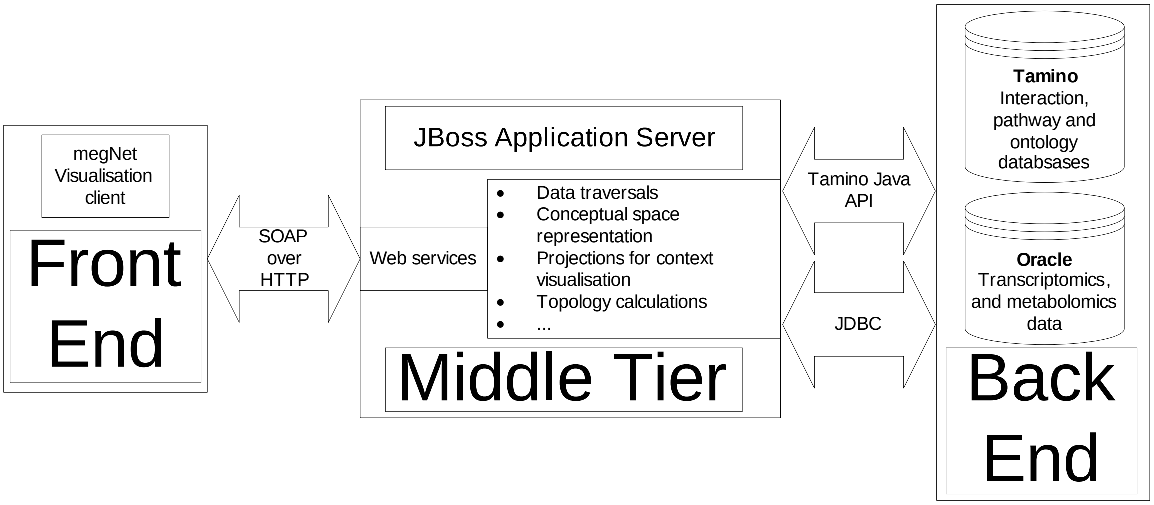


Figure 3.2: Architecture of *MegNet*

Relational databases

Relational database systems are based on relational data model (Codd, 1970) which organizes data as relations or tables and provides a high level query language called Structured Query Language (SQL) for manipulating and querying the relational data (Garcia-Molina et al., 2002). Due to its rigorous mathematical base in Relational Algebra, relational database systems became *de facto* standard in DBMS technology. Oracle (`http://www.oracle.com`) is a famous vendor of a powerful relational database management system.

XML databases

eXtensible Markup Language (XML, `http://www.w3.org/xml`) is a World Wide Web Consortium standard for document markup. XML quickly gained popularity as data exchange format due to its semantic capabilities and its supporting standards and technologies such as DTD, XMLSchema, DOM, SAX, XPath, XQuery and so on (`http://www.w3.org`). Native XML as well as XML enabled database systems have responded to increasing XML data management needs (Chaudhri et al., 2003). Tamino XML Server (`http://www.softwareag.com`) is a powerful native XML data management system. XML is a key technology in biology data management (pp. 291–319 of (Chaudhri et al., 2003)), and commonly used for data exchange based on many XML data standards (Spellman et al., 2002; Hucka et al., 2003; Hermjakob et al., 2004) and for data integration (Achard et al., 2001; Philippi and Köhler, 2004).

3.1.2 Overview of *MegNet*

Our data integration and visualisation system is composed of three layers as depicted in Figure 3.2. Data curated from heterogeneous biological sources including the ontologies constitutes the back-end of the system. The logic and algorithms such as those for the database traversals, network construction, network projections, and integration of gene expression data constitute the middle-tier. The visualisation client communicates with the middle-tier, which in turn communicates with the back-end. Fundamentally the queries that visualisation software enables can be divided into two types: data integration queries and network projection queries. In response to a data integration query sent by the visualisation front-end, the middle-tier queries the relevant databases and performs the integration using database traversals and optionally using additional correlation calculations and sends the results in the form of an integrated network. A network projection request can be made when an integrated network is already constructed, and the system does not use the database back-end in processing these requests. The result of a network projection request is a low-dimensional representation of the network, typically a 2-d plot of network nodes, where the distances between points reflect the similarity of nodes in the network. The similarity measures will be described later in this chapter.

3.1.3 Infrastructure

A detailed description of the infrastructure of the *MegNet* system is included in section *Architectural design* of **Article II**. However, for completeness, it is depicted in Figure 3.2 and a brief summary is given below.

Back-end: The XML data management system Tamino XML server (Software AG) is used for storing XML data. Oracle 10g database server (Oracle Inc.) is used for storing molecular profile data such as gene expression data and metabolomics data. Tamino Java API and Java Database Connectivity (JDBC) enable the communication between the middle-tier and Tamino and Oracle respectively.

Middle-tier: JBoss application server (Redhat Inc.) hosts the middle-tier, which is a set of Java Beans invokable by web services, and sends responses as web services. The communication between middle-tier and the visualisation front-end is handled via SOAP messages.

Front-end: The front-end consists of the user interface. Initially it was implemented in Java, but more recently it is implemented in C#. NET, and the Java client has been deprecated. The client communicates with middle-tier via SOAP messages.

3.1.4 Databases and data curation

Data from various public data sources were collected into our local database systems (Table 3.1, **Article I**, **Article II**). The curation of a public database involves several steps (section 2.3 of **Article I**). Usually every database might need some specialized steps in the curation process (Table 3.1), but the general steps are as follows:

- creating logical schemas which represent the logical structure and physical properties such as indexing to enable efficient queries.
- development of parsers to convert the non-XML data into XML or relational formats or to convert from one XML to another XML format *etc.*

Table 3.1: Databases integrated into our system, along with information on some high level steps employed for integration. Appropriate schemas (XML or relational) were defined for each database. In some cases the logical schemas were readily available *e.g.,* as XML schema definition (XSD) files, in which case they were customised as needed. If the XSD files were not available, they were developed manually. Tamino or Oracle specific physical schema structures including indices for efficient data retrievals were defined based on an analysis of common queries. Tamino mass data loader (or similarly Oracle SQL loader) was used for loading data.

Database	**Curation tasks**
UniProt (Bairoch et al., 2005)	Data for UniProt-SwissProt and UniProt-TrEMBL were available in XML format. Perl parser was written to extract some subset of XML nodes and to format the data for Tamino data loader.
BIND (Bader et al., 2003)	Data was available in XML format. A Java parser was written to format the data for Tamino data loader. In order to link the protein interactors to UniProt entries, the GenInfo identifiers were converted to UniProt accession numbers using SeqHound API (Michalickova et al., 2002)
DIP (Xenarios et al., 2002)	Data was available in XML format. A Perl parser was written to format the data for Tamino data loader.
MINT (Zanzoni et al., 2002)	Data was available in PSI-MI XML format (Hermjakob et al., 2004). A Perl parser was written to format the data for Tamino data loader.
BioGrid (Stark et al., 2006)	Data was available in PSI-MI XML format. A Perl parser was written to format the data for Tamino data loader.
KEGG (Kanehisa et al., 2004)	Pathway data was available in KGML, an XML format (Kanehisa et al., 2004). A Perl parser was written for further Tamino specific formatting. Data from KEGG LIGAND (Goto et al., 2002) database were available as formatted text files. For these data, XML schemas were developed and Perl parsers were written for constructing valid XML documents.
iND750 (Duarte et al., 2004)	Data were available as flat files. Parsers based on libSBML were developed to construct SBML representation (Hucka et al., 2003), and the data pertaining to compound identities were manually annotated with PubChem database identifiers (Wheeler et al., 2007).
TRANSFAC (Matys et al., 2006)	Data were available as formatted text files. XML schemas were developed and Perl parsers were written for constructing valid XML documents.
TRANSPATH (Krull et al., 2006)	Data were available in XML format. Perl parsers were written to extract some subset of XML nodes and for Tamino specific formatting.

GO (Ashburner et al., 2000)	Data were available in an RDF-based format, and correspondly XML schemas were readily available. A Perl script was written for Tamino specific formatting, and for inserting an additional ”Ontology” element which describes whether a term is a “Biological Process”, “Cellular component” or a “Molecular function”. Due to the cyclic dependency of the root elements “go:go” and “rdf:RDF”, an empty schema was first defined in Tamino for “rdf:RDF” element. Then the schema for “go:go” was defined, and finally, the full structure of “rdf:RDF” element was redefined.
GEO (Barrett et al., 2005)	Data was available in simple line oriented text file format called SOFT format. Perl parsers were developed to extract information from the SOFT files. UMLS annotations for meta-data (Butte and Kohane, 2006; Bodenreider, 2004) were incorporated into *maps* database.
T1DBase (Hulbert et al., 2007)	The data were available as MYSQL sqldumps. A MYSQL database was created from the sqldumps, and then the MYSQL data has been exported to ORACLE using Oracle Database Migration Workbench.

Database traversals using schema maps

Biological system description involves various levels of abstraction (Figure 3.1) which include biological molecules (*i.e.,* DNA, RNA, Proteins, Metabolites, and so on), biomolecular interactons (*i.e.,* Protein-protein, Protein-DNA interactions, Transcriptional regulation steps such as a transcription factor binding to the binding site of a gene in order to control its expression, Metabolic reactions), pathways (*i.e.,* cascades or sets of interactions working in concordance in order to perform biological functions), and biological processes (*i.e.,* biological mechanisms involving, for instance, cell-cell, cell-tissue, organ level interactions, and so on). Some of such information can be accessed from the biological databases. Therefore, the integration of such databases allows us to automatically mine that information. In order to build a platform, which allows mining of a variety of such biological interactions, one should fundamentally achieve traversals across the databases providing biological information (Orešič et al., 2004).

Resolving even simple biological relationships that contain a few biological components often requires traversing across multiple databases. The traversals can be achieved by identifying names or identifiers in different databases, which represent the same biological entity. In order to enable traversals in our data integration system, we developed a database called “maps” database, which maps names used for the same entities across multiple databases. This database is populated by parsing information from several databases (Fig. 3B of **Article I**).

Currently this database contains mapping of proteins and gene expression experiments. The protein maps contain information of proteins indexed by Uniprot Swiss-Prot and TrEMBL identifiers (Bairoch et al., 2005). The information was primarily parsed from Uniprot Swiss-Prot and TrEMBL and further populated by supplementary information parsed from BIND database (Bader et al., 2003), KEGG Pathways (Kanehisa et al., 2006), and yeast metabolism models iND750 (Duarte et al., 2004) and YMN1_0 (Herrgård et al., 2008). Maps for gene expression experiments contain experimental meta descriptions and annotations from Unified Medical Language System (UMLS) where available (Butte and Kohane, 2006; Bodenreider, 2004).

As an example to illustrate how the data in *maps* database is compiled, we explain how protein entries in this database are linked to GenInfo identifiers provided in BIND molecular interaction database (Bader et al., 2003). In protein-protein interaction databases such as MINT and DIP, protein nodes are indexed by Uniprot identifiers (Zanzoni et al., 2002; Chatr-aryamontri et al., 2007; Xenarios et al., 2000). But in BIND, interactions are indexed by GenInfo sequence identifiers (GI's) (Bader et al., 2003). We obtained mapping between the GI's and Uniprot identifiers in each BIND interaction using DBXREF table that provides all external references made in BIND and SeqHound Perl API (Michalickova et al., 2002) as follows. We first extracted the list of unique UniProt accession numbers from the DBXREF table. Then, for each Uniprot accession number in this list, we did the following. We first found the corresponding primary GI using `SHoundFindAcc` function of SeqHound Perl API. Next, we retrieved the list of GI's that have the exact sequence as that of the primary GI using `SHoundRedundantGroup` function of the API. We finally stored mapping between all the GI's whose taxonomy id (as found by `SHoundTaxIDFromGi` function of the API) matched that of the primary GI. Finally, all GI's associated with each UniProt accession were added as external database links for the corresponding protein entry, indexed by the UniProt accession number, in *maps* database.

The data traversals use protein nodes as *central nodes* for constructing integrated networks. More specifically, in order to construct an integrated network consisting of various types of edges in it, those edge types are joined based on common protein nodes they share. As an example, consider traversing from a metabolic reaction to a protein-protein interaction. Such traversal could bring information about the protein-protein interaction partners, when exist, of the enzyme which catalyzes the metabolic reaction. It can be typically achieved by translating the Enzyme Classification (EC) number of the enzyme to the corresponding protein identifier (Uniprot Id), and then searching protein-protein interaction databases for any possible interactions partners (Fig 3 of **Article I**). Similarly, for linking correlation networks based on gene expression measurements with interaction networks we find out identifiers of the proteins which are expressed by the genes (indexed by EMBL sequence identifiers) under consideration.

In order to achieve data traversals, the *maps* database is designed to contain identifiers and names of protein entities from multiple databases. Conversions from other types of identifiers, *e.g.,* EC enzyme identifiers to Uniprot protein identifiers, and linking of *e.g.,* EMBL gene identifiers to Uniprot protein identifiers are obtained from *maps* database.

3.1.5 Integration of Gene expression data

We achieved integration of gene expression data with interaction networks in other ways (sections 3.2, 4.2). At this point, however, we limit the discussion to network visualisation based integration approach. The visualisation approach is based on correlation networks. For selected biological samples, possibly from one or more gene expression studies, the integration approach will be explained below.

Gene co-expression across multiple experimental conditions and multiple organisms indicates strong functional relationship, and hence is a powerful tool for elucidation of gene function (Stuart et al., 2003). Gene Expression Omnibus (GEO) (Edgar et al., 2002; Barrett et al., 2005) stores high-throughput gene expression data from many organisms and a huge variety of biological conditions.

In *MegNet* we allow the users to assess co-expression of genes from any choice of combinations of data sets and integrate the information with pathway and interaction data.

Gene expression data preprocessing

Many variations introduced at different stages of microarray experiments blurr the real biological variation and microarray data preprocessing tries to remove the non-biological variations from the data (Speed, 2003; Quackenbush, 2002; Bolstad et al., 2004). Within each GEO data set, the statistical preprocessing of the data such as background correction and normalisation are performed in a consistent fashion (Barrett et al., 2005), thereby removing the experiment-specific non-biological variations. But in order to enable simultaneous mining of gene expression data from multiple studies together, we need further normalisation so that different data sets can be combined. For example, in building ONCOMINE system to mine a large number of diverse cancer microarray data sets, Rhodes et al. (2004) applied median centering and standard deviation scaling per microarray. In an across-laboratory reproducibility study of microarrays, Irizarry et al. (2005) indicated that studying relative expression values instead of absolute expression values is a simple solution to remove the probe-specific effects in particular experiments.

In the data collected from GEO, there are two major types of gene expression microarray datasets: single channel (*i.e.,* intensity based) microarrays (Lockhart et al., 1996) such as Affymetrix oligonucleotide micoarray data, and dual channel (*i.e.,* two colour) microarrays (Schena et al., 1995) such as cDNA microarray data. Therefore, the cross-platform comparability problem reduces to applying normalisation so that both these dataset types can be analysed in an identical fashion. Since GEO data provides log2 ratio between individual channel intensities (*i.e.,* between case and control) for the dual channel arrays, we similarly normalise the single channel data of case samples with control samples from the same study, where the selection of the case and control samples is upto the user. To be more precise, for each case sample, we compute the log2 ratio of the gene expression intensity measurement versus the average intensity of control samples from the same GDS dataset. After this transformation, single channel data and dual channel data can be analyzed in identical fashion as well as simultaneously. In order to allow data sets coming from different studies to be combined, we scale each microarray to unit standard deviation. Finally, when multiple data sets are queried, only the genes common to all microarray platforms are used for the analysis.

Statistical hypothesis testing

Statistical hypothesis testing (Box et al., 1969; Montgomery, 1983) is a framework that allows us to answer particular questions related to one or more populations on the basis of samples randomly drawn from those populations. A statistical hypothesis is an assumption about the probability distribution of a population. Hypothesis testing generally involves the following steps

Formulation of null hypothesis (H_0): The null hypothesis is the hypothesis that the results observed in a study (*e.g.,* difference between treatment group and control group) are purely by chance. The null hypothesis would be rejected if data does not provide enough evidence to its truth; otherwise, we fail to reject it. An alternative hypothesis (H_a) is complementary to the null hypothesis, and is effectively favoured when the null hypothesis is rejected.

Calculating a test statistic: The test statistic is a measure of the size of the "effect" relevant to our test.

Calculating the P-value: The significance probability or P-value is the probability of getting the data at least as extreme as observed if the null hypothesis were true.

Rejecting or failing to reject the null hypothesis: The level of significance (α) is the maximum probability with which we are willing to falsely reject the null hypothesis. The null hypothesis would be rejected if the significance probability (p-value) is less than the level of significance (α); the evidence to reject the null hypothesis is considered inadequate otherwise.

Parametric tests make particular assumptions on the properties of the populations. When the parametric assumptions are not satisfied, *non-parametric tests* may be used instead. The advent of powerful computers allows a new approach called a *permutation test* (Moore and McCabe, 2005) for performing a *non-parametric test.* A *permutation test* employs resampling to estimate the significance of the test statistic. Permutation tests are often more robust than formula based non-parametric or parametric tests. They are applicable even though the parametric assumptions are not satisfied, as long as the resampling is done in a way that is consistent with the null hypothesis (Moore and McCabe, 2005).

Multiple hypothesis testing

In the context of hypothesis testing, two types of errors are possible: Type I error or Type II error. Rejecting a true null hypothesis is called the Type I error. Failing to reject a false null hypothesis is called the Type II error. When multiple hypotheses are tested at a specified Type I error probability (α) for each test, the chance of committing at least one Type I error increases sharply with the number of hypotheses, and such phenomenon is referred to as multiple comparison (or testing) problem. In order to address the multiple comparison problem, one needs to define an appropriate Type I error rate and devise powerful multiple testing procedures that control this error rate (Shaffer, 1995; Dudoit et al., 2003). Controlling the false discovery rate (FDR) is one such approach, in which, the FDR—the expected proportion of falsely rejected null hypotheses (Benjamini and Hochberg, 1995)—is controlled. Procedure for controlling FDR proposed by Benjamini and Hochberg (1995) is as follows. Sort the p-values of the test in ascending order, and denote the sorted order as

$$p_1 \leq p_2 \leq p_3 \leq \ldots p_m.$$

Represent the corresponding m hypothesis with the same indices

$$H_1, H_2, H_3, \ldots, H_m.$$

For controlling the FDR at level α, define

$$j_0 = max\{j : p_j \leq \frac{j}{m} \cdot \alpha\}$$

and reject hypothesis

$$H_i, \forall i \in 1, 2, 3, \ldots, j_0.$$

Other classical Type I error rates controlled in multiple comparison approaches include Per Family Error Rate (PFER)—expected number of type I errors, and Family Wise Error Rate (FWER)—the probability of at least one type I error (Shaffer, 1995), both of which tend to impose more strict control over the Type I errors, but at the expense of larger number of Type II errors. Therefore, controlling the FDR is common choice in exploratory analysis (Dudoit et al., 2003).

Correlation

Correlation between two variables X and Y is the extent to which their values vary together systematically (Box et al., 1969). A correlation coefficient is a measure to the extent of correlation between two variables. Product-moment correlation coefficient or pearson correlation coefficient, which is calculated from given samples of X and Y as the ratio between their covariance and the product of their standard deviations, indicates the extent to which relationship between the variables X and Y is linear. Pearson correlation coefficient obtained from samples (r) is an estimate of the true correlation (ρ) of the bivariate normal population (X, Y). The null hypothesis that $\rho = 0$ can be tested with the statistic

$$t = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}}$$

which follows Student's distribution with $(n-2)$ degrees of freedom (Box et al., 1969). The p-value of this test is the probability that pearson correlation coefficient of the sample (r) would be at least as extreme as observed if the null hypothesis is true (*i.e.,* if $\rho = 0$).

Co-expression network

MegNet allows studying co-expression of genes from multiple GEO data sets emanating from different labs, different studies, and possibly using diverse microarray platforms. Genes that are common to all microarray platforms in the query are identified and used for the combined analysis using correlation via Pearson correlation coefficient. For visualizing gene expression data, we compute correlation and p-value between every pair of genes in the selected datasets. We then apply FDR method of Benjamini and Hochberg (1995) to account for the multiple hypothesis testing. A correlation network is then defined as the collection of edges representing statistically significant correlation between the pairs of gene nodes they connect. Edges sharing common genes are joined at the common gene nodes.

Finally data traversals can be also combined with co-expression network construction in order to link the correlation network to molecular interaction networks, pathways and ontologies. For example, *MegNet* system may be used to look for interaction neighbourhood of entities on the correlation network. But more generally, it is possible to combine interaction network construction parameters (*e.g.,* certain proteins, pathways or ontologies) and co-expression network construction parameters (*e.g.,* some particular selection of gene expression data sets) in the same query, and then let the data traversals automatically discover the cross-talk between these different levels wherever possible (see *e.g.,* section 4.1.4).

3.1.6 Context-sensitive data mining

Conceptual spaces

Gärdenfors (2000) advocates conceptual spaces as a representation paradigm to modelling human cognition and as complementary to the Symbolic and Associationist levels of representation used in cognitive science. He argues that these three levels are representations at different levels of resolution. In conceptual spaces, concepts are represented as high-dimensional geometric spaces over a variety of quality dimensions with geometric or topological structure for one or more domains. A domain can be modelled as a set of integral dimensions *i.e.,* dimensions which need not necessarily be completely independent. For example, an apple might be represented as a multi-dimensional space where the dimensions could be "colour", "taste", "shape" and so on. Thus, while an "apple" is simply a token (or a text string) at the Symbolic level, it has a rich underlying geometric description at the conceptual level.

There are two approaches to explaining meanings: *Realist* and *Cognitive* semantics. According to *Realist semantics*, meanings of expressions are independent of the individuals involved in communication; meanings are in the world. According to *Cognitive semantics*, meanings are mental entities;

different people have different conceptualisations of the world. The realist semantics fail to deal with the fact that different people have different approaches to learn and that meanings of concepts often change over time and between contexts. Context plays an important role in human learning or understanding of new concepts. In terms of conceptual spaces, different conceptualisations correspond to considering different subsets or alternative choices of quality dimensions. The dynamic nature of the meaning, *i.e.,* changes in meaning of concepts with respect to time and context, can be represented by giving different saliencies (*i.e.,* weights) to dimensions and domains.

Context-sensitivity in conceptual spaces

Context-sensitivity can be achieved in conceptual spaces by specifying weights to the quality dimensions and recomputing the similarities (Gärdenfors, 2000). However, how to appropriately assign weights to the dimensions according to the context is not necessarily trivial.

In building an information inference application, Song and Bruza (2003) employed a heuristic based on query term frequency in text corpus to automatically find out the dominant concept in a concept combination. The *concept combination heuristic* was then applied to emphasise quality properties shared by the concepts in the concept combination and to increase the weights of properties of the dominant concept by rescaling (Song and Bruza, 2003). Raubal (2004) in their way-finding application used different sets of empirical weights to represent the users preferences on facades according the day and night.

However, the true power of conceptual spaces representation is in allowing each user to have one's own conceptualisation of the world. Therefore, in an interactive system, the user must be given full control of weight assignment. This is especially useful for facilitating exploratory data mining. Meanwhile, the system may offer some predefined or default contexts defined by some particular choices of weights in order to facilitate queries from users with little experience.

Using some specific experimental context of interest, if available, maybe a useful alternative. For example, a gene expression experiment data maybe used to assign weights based on some measure such as correlation coefficient (see *e.g.,* section 4.1.5).

Conceptual representation of *MegNet* networks

An integrated network constructed using database traversals as explained earlier corresponds to symbolic representation of biological relationships obtained using semantic web technologies. Therefore representing the information in the networks as a conceptual space makes it possible to mine the data in a context-sensitive manner, or in other words, to facilitate *Cognitive semantics.*

In this direction, we represent every *MegNet* integrated network as a high-dimensional space (Figure 3.3) whose dimension is defined by the number of nodes in the network (**Article IV**). Every node in the network is a vector in this space. A vector v representing a node p contains weighted shortest path distances from node p to all other nodes in the network. These weights are the subject of discussion in the next paragraph. For now, it is enough to note that, by default, the weights for all interactions are set to 1. Therefore, the default representation contains shortest path length for each coordinate.

As discussed earlier, context-sensitivity in a conceptual space can be achieved by specificying one's own conceptualisation in terms of weights on quality dimensions. But translating one's conceptualisation to weight assignment may not always be trivial. In fact, finding the best weights may need exploration. Therefore, to facilitate interactive exploration, as a pragmatic alternative, we give the full control of the weights to the user, and only define the default context (or system's view of the world), in which we assign a unit weight (*i.e.,* weight of 1) to each interaction type.

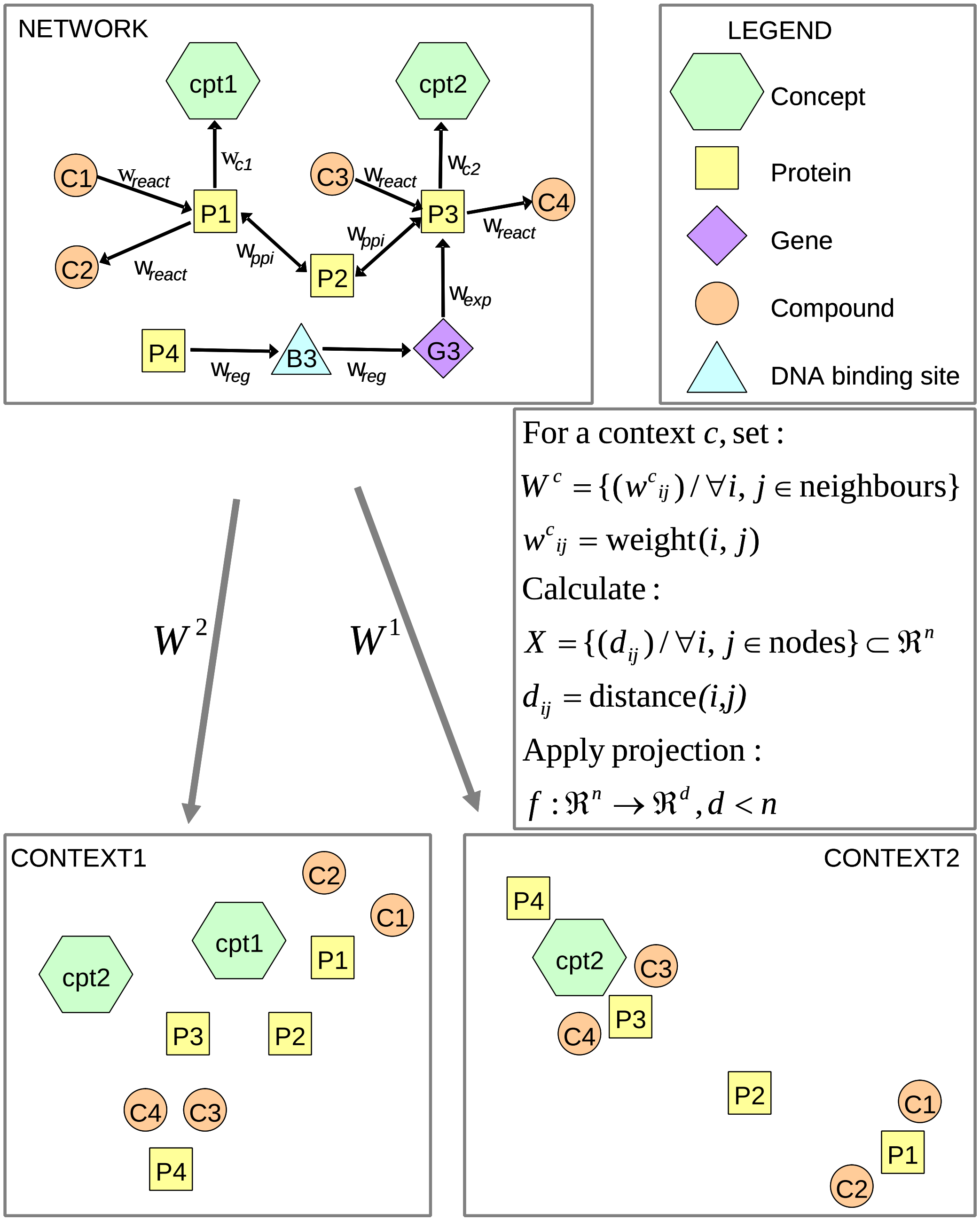


Figure 3.3: Conceptual space representation and context visualisation of *MegNet* networks. A context is defined by a specific choice of weight assignment to the edges. Every network node is a vector in a high-dimensional space where each coordinate represents the weighted shortest distance of this node from a particular node in the network. Dimensionality reduction methods enable us to visualise this space in a low-dimensional (typicaly 2-d) plot.

As the integrated network consists of heterogeneous edge types, translating the user's context specification (*i.e.,* preferences on weighting the quality dimensions) to the assignment of weights to individual edges (or weight normalisation) is not obvious, because the underlying biological meaning of the edges has to be preserved. For example, a metabolic reaction consists of two edge units: the first edge connecting substrate to enzyme, and the second connecting enzyme to the product. Therefore, assigning a weight of 1 unit to a metabolic reaction is, for instance, achieved by assigning a weight of 0.5 to each of these two edges. For each datatype incorporated in our database system, we define the standardisation for the distances in *MegNet*, so that the user is not required to know the underlying representation of the data while specifying weights to interactions.

Visualisation of a context

Typically, *MegNet* integrated networks are very complex in the sense that they contain thousands of nodes and edges. Hence, understanding the conceptual space encoded in such a network under a particular context is a formidable task. Visualising the semantic similarities encoded into the conceptual representation allows us to visualise a particular context specified by the user. Therefore, as a solution to the visualisation problem, we apply dimensionality reduction techniques (Carreira-Perpiñan, 1997) and visualise the distances in a lower-dimensional space (typically 2-d) (Figure 3.3). Dimensionality reduction is most often possible, because the original high-dimensional representation of the data often contains redundancies and dependencies between the variables. Under the assumption that the data lie on a manifold whose dimension is smaller than the dimension of the embedding space, the dimension reduction can be achieved by the construction of a continuous mapping between the embedding space and the unknown manifold space. This mapping has to be inversible in order to project and reconstruct the original data with minimal error (Lee et al., 2004). The true dimension of the unknown manifold is known as the intrinsic dimension of the data. The intrinsic dimension needs to be provided as an input to most dimension reduction techniques to build a correct mapping *i.e.,* to avoid over or under fitting. Thus, finding the correct value for the intrinsic dimension might be an exploratory undertaking. However, 2-dimensional plots are the easiest and the most common choice for visualisation. A variety of dimension reduction techniques exist, and are usually utilised for a variety of purposes such as visualisation, data compression, and variable selection (Carreira-Perpiñan, 1997). As we are particularly interested in the visualisation of the semantic distances, our focus is on dimension reduction techniques that preserve the pairwise distances of the data space in the projection space. Therefore, we implemented three non-linear dimensionality reduction methods in *MegNet*, namely, Sammons mapping (Sammon, 1969), Curvilinear component analysis (CCA) (Demartines and Hérault, 1997), and Curvilinear distance analysis (CDA) (Lee et al., 2000, 2002, 2004). The aim of these projection methods is to represent the high-dimensional vectors in a low-dimensional space in such a way that the distances are preserved by the mapping. Each method achieves this objective by minimising a cost function. Technical details such as the analytical formulae of the cost functions, the computational complexities of these algorithms, and our implementation of these ideas are discussed in detail in **Article II** and **Article IV**. Some important properties of these methods are given here for completeness.

Sammon's mapping (Sammon, 1969) estimates the configuration of entities in the projected space using steepest gradient descent algorithm on a cost function that is based on the interpoint distances between the entities in the original space and the discrepancies introduced by the dimensionality reduction mapping (see section 3.3.1 of **Article II**). In this way, the visual configuration approximates the original relationships in the complex networks. Sammon's mapping puts higher emphasis on preserving smaller distances. Curvilinear component analysis (CCA) (Demartines and Hérault, 1997) attempts to preserve the topology by first favouring shorter distances, and then the longer distances. CCA uses stochastic gradient descent. Curvilinear Distance analysis (CDA) (Lee et al., 2000, 2002, 2004) maps entities in higher-dimensional space into lower-dimensional space in such a way that the curvilinear distances, as opposed to Euclidean distances, between points in high-dimensional space are preserved. Curvilinear distance is the shortest path distance in the centroid graph in the high-dimensional space. CDA uses stochastic gradient descent algorithm to minimise the projection error.

3.2 Dynamic topology of integrated networks

High-throughput functional genomics techniques such as gene expression microarrays (Schena et al., 1998; Lockhart and Winzeler, 2000), mass spectrometry based proteomics (Aebersold and Mann, 2003) and metabolomics (Goodacre et al., 2004; Orešič et al., 2006) allow us to measure the state of a biological system in terms of molecular abundances or concentrations. Databases describing signalling (Krull et al., 2006), metabolic pathways (Kanehisa et al., 2004; Goto et al., 2002) and biological functions (Ashburner et al., 2000) provide reference information on cellular processes. One important goal of developing our data integration platform is to enable interpretation of new experimental data in the context of the established pathway knowledge.

3.2.1 Transcriptomics

Transcriptome is the collection of messenger RNA (mRNA) molecules or “transcripts” in a cell or a tissue. Transcriptomics is the discipline which deals with the large scale study of the transcriptome under selected conditions. DNA microarray technology (Schena et al., 1995; Lockhart et al., 1996), capable of measuring activity of thousands of transcripts simultaneously, has served as the platform of choice for transcriptomics.

3.2.2 Metabolomics

The word metabolome represents the whole collection of metabolites in an organism. Metabolomics is defined as the comprehensive study of the whole metabolome under particular conditions (Fiehn, 2001; Goodacre et al., 2004). It embodies global study of all metabolites, their dynamics, composition, interactions, responses to interventions or environmental changes (Orešič et al., 2006). Owing to the fact that small changes in the activities of individual enzymes can lead to large changes in metabolite concentrations, metabolomics is a more sensitive tool than transcriptomics and proteomics for studying complex diseases (Orešič et al., 2006) and the regulatory roles of nutrition in human health (Gibney et al., 2005). But due to the huge diversity among metabolites, different technologies are required for studying different classes of metabolites, yet it is impossible to study the whole metabolome (Orešič et al., 2006). An interesting subfield of metabolomics is lipidomics, a discipline which deals with global study of lipids. Lipids are highly diverse molecules which play crucial roles in cellular energy storage, structure and signalling (Orešič et al., 2008; Seppänen-Laakso and Orešič, 2009).

3.2.3 Pathway analysis

A challenge in the analysis of data arising from functional genomics experiments is to understand the results in the context of established information such as biological processes, metabolic pathways and so on. Understanding what biological processes are significantly modulated under a particular intervention is often more informative than finding a set of individual molecules that are differentially regulated (Curtis et al., 2005).

Recently a set of techniques broadly known as Pathway analysis methods have been developed to achieve this goal (Curtis et al., 2005). A most prominent technique that enables pathway analysis of genome-wide gene expression data is Gene Set Enrichment Analysis (GSEA) (Mootha et al., 2003; Subramanian et al., 2005). By determining which gene sets (*i.e.,* groups of genes that share common biological function, chromosomal location or regulation) are significantly enriched among the significantly modulated genes in a gene expression study, GSEA not only offers an easy interpretation of the data, but also accounts for the fact that single-gene analysis (*i.e.,* analysis of differential expression using a statistical test) may miss important "group or pathway effects". By definition, a gene set represents the set of genes that act in concert to "perform" a biological function. A moderate but concordant change in a gene set therefore may be more important than a very high change in a single gene in the set. This is called the "group effect or pathway effect". Moreover, Subramanian and colleagues (Subramanian et al., 2005) also showed, by applying GSEA on three different lung cancer data sets, that the pathway level changes are more reproducible than the individual gene markers. Furthermore, GSEA is not only limited to transcriptomics data analysis. Subramanian et al. (2005) said, "GSEA can clearly be applied to other data sets such as serum proteomics data, genotyping information, or metabolite profiles". Recently GSEA algorithm has also been adapted for the data analysis in genome-wide association studies (Wang et al., 2007; Holden et al., 2008).

Numerous conceptually similar but competing approaches to GSEA exist, which employ alternative enrichment statistic and permutation testing scheme, seeking to improve the statistical power of the pathway analysis (Tian et al., 2005; Efron and Tibshirani, 2006). Moreover, high quality pathway databases would enhance the usefulness of pathway analysis. Multiple proprietary pathway databases and pathway analyses software also exist (`http://www.ingenuity.com`, `http://www.ariadnegenomics.com`, `http://www.genego.com`).

Finally, several studies reported that gene level biomarkers are not reproducible (Tan et al., 2003; Michiels et al., 2005; Ein-Dor et al., 2005). Despite this poor reproducibility, however, pathway level changes are more consistent (Subramanian et al., 2005; Zhang et al., 2008). Thus, pathway level analysis is a more promising tool for identifying the disease mechanisms, and adaptive physiological compensatory responses (Curtis et al., 2005).

3.2.4 Topology of biological networks

Topological properties of a variety of biological networks such as protein-protein interaction and metabolic networks have been extensively studied by many researchers (Barabási and Oltvai, 2004). Due to intriguing similarities such as scale-free topology between the topological properties of biological networks and other networks such as social networks, scientific collaboration networks and so on, Network biology has been an active area of research. Besides analysing the structures of static networks, studying the network dynamics is of interest as it may enable elucidation of dynamic design principles. Kharchenko et al. (2005) studied the expression dynamics of a metabolic network and discovered relationships between pairwise distances on the metabolic network and the co-expression of genes. Luscombe et al. (2004) studied regulatory network dynamics by integrating transcriptional regulatory information and gene-expression data, and showed that, in response to diverse stimuli, transcription factors alter their interactions to varying degrees causing large topological changes in the regulatory network.

In order to study the topological dynamics of integrated networks at the level of functional modules and to facilitate analysis and interpretation of molecular profile data such as gene expression data via integrated network connectivity, we have established a method called Topological Enrichment

Analysis of Functional Subnetworks (TEAFS). This method benefits from genome wide integrated networks constructed using *MegNet* and genome wide gene expression profiles. **Article V** reports an application of this method for studying dynamic response of oxidative stress, the results of which are presented in section 4.2.

3.2.5 Integration of transcriptomics and interactomics data

In a recent study, subnetwork markers identified by integration of protein-protein interaction networks and gene expression data (Chuang et al., 2007) were found to be more reproducible than individual gene markers found by differential gene expression analysis and they achieved higher classification accuracy in discriminating metastatic versus non-metastatic tumours. Genes with known breast cancer mutations are typically not significantly differentially expressed, but were found to play central roles in protein networks by interconnecting many differentially expressed genes. This demonstrates the value of studying the gene regulation at the level of interaction network connectivity.

Although gene set enrichment analysis (GSEA) (Subramanian et al., 2005) and the related methods account for the subtle but coordinated patterns of gene expression changes, they do not take the connectivity of the system into account. For example, loss of a central node from the interaction network representing the biological process may lead to compensatory rise in the expression of multiple genes involved in the biological process simultaneously, but the loss of connectivity would render the network dysfunctional. However, due to the simultaneous rise in the expression of multiple genes, GSEA would report this dysfunctional pathway to be significantly up-regulated.

Recently a method called Gene Network Enrichment Analysis (GNEA) has been developed (Liu et al., 2007). This method takes the connectivity of proteins in protein-protein interaction networks into account in order to find pathways that are consistently affected across multiple interventions or models related, for instance, to a particular human disease, using gene expression data. Thus, the philosophy of this method is not to identify biological processes perturbed under a particular intervention but rather to discover which biological processes are transcriptionally altered across diverse tissue types in the context of a disease.

Outline of TEAFS

Topological Enrichment Analysis of Functional Subnetworks (TEAFS) facilitates understanding of how a specific biological intervention modulates biological functions. Below we provide a general outline of the TEAFS method. More specific details of TEAFS for studying dynamically most changing modules can be found in **Article V**.

Construction of an integrated network: TEAFS uses connectivity of protein nodes in an interaction network in terms of various topological measures (Albert and Barabási, 2002) to compute the topological changes of subnetworks representing functional modules. The changes in the topology of the subnetworks indicate the the changes in activity of functional modules. Thus an interaction network is an input to this method and forms the starting point for this analysis. Any genome wide interaction network of the organism of interest, such as a protein-protein interaction network can be used for this purpose, but we take the unique advantage of the *MegNet* integrated networks.

Gene detection: The second input to TEAFS is the gene expression data that is intended to be analysed. This data is used to compute whether a protein is “present” (P) or “absent” (A) based on whether or not the transcript encoding the protein shows sufficiently high expression level. For probeset detection in Affymetrix Oligonucleotide arrays, MAS5 software (`http://www.affymetrix.com`) uses Wilcoxon signed rank test for perfect match (PM) and mismatch (MM) probes, and "calls" the probeset "present", "absent" or "marginal" (M) based on the p-value (m Liu et al., 2002). More recently, an alternative gene detection method called PANP was introduced for computing P/A calls for oligonucleotide arrays (Warren et al., 2007), which uses only the PM probes for Affymetrix chips. For spotted DNA arrays, if the hybridization signal intensity of a spot is above the background intensity distribution, the corresponding gene may be considered present. Luscombe et al. (2004) applied one such criterion for computing the presence or absence of transcription factors in order to reconstruct transcriptional regulatory networks. For each sample in the selected gene expression study, we compute the P/A corresponding to all transcripts analysed on the chip. For studying the effects of yeast oxidative stress, we employed the P/A criteria similar to Luscombe et al. (2004) (**Article V**).

Reconstruction of networks: We identify the proteins encoded by the transcripts, for instance by translating the accession numbers, and thus infer which proteins in the network are present and absent in each condition. For each sample in the gene expression study, we construct the condition specific network by removing the absent protein nodes and the edges incident on them from the original network.

Identification of Functional Subnetworks: Functional subnetworks in the integrated network can be identified by using the established knowledge of the functional association of proteins. For example, by identifying which proteins in the network are associated with a particular Gene Ontology term (Ashburner et al., 2000), we can identify the functional subnetwork consisting of all these proteins and their neighbouring interactions. Other functional categories such as metabolic pathways (Kanehisa et al., 2006) may also be used similarly.

Computation of Topological Measures: In a directed network, the number of edges coming into a particular node is called the "in degree" of the node, and the number of edges going out of the node its "out degree". Two nodes connected to each other by an edge are called neighbours. Clustering coefficient of a node is the ratio between the number of edges between its neighbours and the maximum number of theoretically possible edges between those neighbours. A topological measure (i.e. in degree, out degree or clustering coefficient) of functional subnetwork in an integrated network is the average of the topological measure of all nodes that are members of the subnetwork (Albert and Barabási, 2002).

Enrichment score: Enrichment score or enrichment statistic assigned to a functional subnetwork is the quantity of interest that we compute based on what questions we want to answer. In order to find out the most changing subnetworks during a time course following a particular intervention, an appropriate enrichment score would represent the degree of topological change the subnetwork is subjected to over time by the underlying biological intervention. In order to find out the most changing functional subnetworks during the time course following oxidative stress in yeast we employed standard deviation of topological measure as the enrichment score (**Article V**).

Computation of Statistical Significance: To compute the statistical significance of the enrichment score, a null distribution for the enrichment score is calculated, based on the random P/A model of proteins in the yeast oxidative stress study (**Article V**). A permutation test (Moore and McCabe, 2005) was devised for this.

Chapter 4

Results

4.1 *MegNet* based data mining

According to Hand et al. (2001), "*Data mining* is the analysis of (often large) observational data sets to find unsuspected relationships and to summarise the data in novel ways that are both understandable and useful to the data owner", and "the relationships and summaries derived through a data mining exercise are referred to as *models* or *patterns*". *MegNet* data integration system facilitates mining of biological data and hence exploration of some useful patterns, novel relationships between different biological entities from the data, and may provide novel insights into protein functions and context-specific biological functions.

Here, we summarise some *MegNet* data mining examples that were published in **Article I**, **Article II**, and **Article III**. The examples provided in sections 4.1.1 and 4.1.2 describe mining based solely on data traversals. Section 4.1.3 presents a case of a conceptual space representation with unit weights. Section 4.1.4 presents the case of integration between co-expression network with interactions and ontologies. Finally, section 4.1.5 presents context-sensitive mining of interaction and ontology information using the context of the gene expression experiment.

4.1.1 Integrated pathway retrieval

When attempting to model real biological phenomena one needs to understand the cross-talk across different levels of biological organisation, for instance, between metabolic pathways and cell signaling (Papin and Palsson, 2004). *MegNet* models the cross-talk across different levels through database traversals.

As an example, we queried the following *Saccharomyces cerevisiae* metabolic pathways: Glycolysis / Gluconeogenesis, Pentose phosphate pathway, and TCA cycle. More specifically, the query processing would include retrieval of primary components (*i.e.,* enzymes and compounds) of the metabolic pathways from KEGG (Kanehisa and Goto, 2000), and then database traversals to search protein-protein interaction databases BIND (Bader et al., 2003) and MINT (Zanzoni et al., 2002) for interactions of the enzymes with the nearest neighbor proteins.

The resulting networks showed surprisingly high level of connectivity across different stages of linear metabolic pathways via protein-protein interactions (Fig. 5 of **Article I**). Specifically, two enzymes from the glycolysis pathway: phosphoglycerate kinase (PGK; EC: 2.7.2.3) and acetate-CoA ligase (ACS; EC: 6.2.1.1) appeared to aggregate with SRB2 (Uniprot accession: P34162) via protein-protein interactions.

ACS catalyzes formation of acetyl-CoA from acetate, which is a starting point in the TCA cycle, while PGK catalyzes acetylation of 3-phospho-d-glycerate, which is a part of the second phase of glycolysis. SRB2 is involved in transcriptional initiation. This could mean that PGK and ACS, enzymes at two different stages of glycolysis, are co-regulated. Thus, our results point towards a testable hypothesis (**Article I**).

4.1.2 Protein neighbourhood search

Assignment of protein function is a non-trivial task owing to the fact that the same protein may be involved in different biological processes, depending on the state of the biological system and protein localisation (Camon et al., 2004). *MegNet* allows visualisation of interaction neighbourhood of a protein, *i.e.,* the entities in the network close to the protein, which may potentially provide insights into the function and mode of action of a protein. These entities include molecules, genes or more complex concepts.

As an example, we searched the neighbourhood of mannose-6-phosphate isomerase for *Saccharomyces cerevisiae* (PMI40; UniProt Id: P29952), which catalyzes the conversion between fructose-6-phosphate and mannose-6-phosphate and thus connects glycolysis with the cell wall synthesis in *S. cerevisiae.* The search involved concurrent retrieval of relationships for the following databases: UniProt (Bairoch et al., 2005), KEGG (Kanehisa and Goto, 2000), BIND (Bader et al., 2003), MINT (Zanzoni et al., 2002) and GO (Ashburner et al., 2000).

Figure 6 of **Article I** shows the neighbourhood of PMI40. The zoomed-in window shows one region of potential interest, which includes protein-protein interactions between the PMI40 and NUP100 (UniProt Id: Q02629), a subunit of the nuclear pore complex, as well as between NUP100 and alpha-1,6-mannosyltransferase (MNN10; UniProt Id: P50108). Also both PMI40 and MNN10 are involved in cell wall mannoprotein synthesis (GO:0000032).

PMI40 is a gate between cell wall synthesis and glycolysis, *i.e.,* cell decision point between growth or energy production, and MNN10 is a part of the protein complex in mannoprotein synthesis toward the end of the cell wall biosynthesis pathways. Examination of interaction entries (BIND id's: 137955 and 137823) suggests that NUP100 protein, which is a part of nuclear pore complex, binds to the PMI40 and MNN10 open reading frames (Casolari et al., 2004). This and other evidence by Casolari et al. (2004) provide support for the gene gating hypothesis, which suggests that the interaction of the nuclear pore complex with different genes might serve as a level of gene regulation (Blobel, 1985). Thus, again, *MegNet* based mining leads to a testable hypothesis that PMI40 and MNN10 are co-regulated in relation to cell decision-making between energy production versus growth.

4.1.3 Context-dependent protein function visualization

In order to suggest protein function annotations to a set of human proteins related to maintenance of energy homeostasis and specific G-protein coupled receptors (GPCRs) that are not yet well characterized—PPAR-γ, PPAR-α, PGC1-α, SREBP2, GPR40, GPR41, and GPR43—we searched their neighbourhood using protein-protein interaction databases BIND (Bader et al., 2003), MINT (Zanzoni et al., 2002), DIP (Xenarios et al., 2000), metabolic pathway database KEGG (Kanehisa and Goto, 2000), transcriptional regulation database Transfac (Matys et al., 2006) and Gene Onology (Ashburner et al., 2000) databases. The resulting network contained three isolated subnetworks, with all three GPCRs jointly in one subnetwork (Fig. 1 of **Article III**). While some of the well known relationships were revealed in the largest subnetwork, the results of the query have not facilitated characterization of poorly annotated proteins such as GPR40, GPR41, and GPR43.

We then formulated the conceptual space representation of the data, by assigning unit weights to all the edges in the network and visualised the resulting similarity strucure using curvilinear distance analysis (CDA) (Lee et al., 2004) (Fig. 2 of **Article III**). According to this figure, in the underlying conceptual space, PPAR-γ and GPR41 are closely associated with response to nutrients (GO:0007584). PPAR-γ (UniProt id: P37231) is annotated in UniProt as "Receptor that binds peroxisome proliferators such as hypolipidemic drugs and fatty acids. Once activated by a ligand, the receptor binds to a promoter element in the gene for acyl-CoA oxidase and activates its transcription. It therefore controls the peroxisomal beta-oxidation pathway of fatty acids. Key regulator of adipocyte differentiation and glucose homeostasis". Thus, within the specific context of relationship to a specific GPCR, the context-based mining approach improves this annotation by indicating that PPAR-γ and GPR41 are closely associated with response to nutrients, as also supported by recent research (Xiong et al., 2004).

4.1.4 Integrated co-expression and interaction networks

As an illustration of combining gene expression data with the existing interactions, pathways and ontologies, we have utilised gene expression data from mouse congenic strains in a study related to Type 1 Diabetes (Eaves et al., 2002). Type 1 Diabetes (T1D) is an autoimmune disease caused by destruction of pancreatic beta cells. Non Obese Diabetic (NOD) mouse (Makino et al., 1980) is a model of autoimmune diseases; it develops spontaneous autoimmune diabetes, which shares many similarities with T1D in humans. Hence, NOD mouse serves as an excellent animal model for T1D (Anderson and Bluestone, 2005).

"A congenic strain is a strain identical or almost identical to a standard inbred strain except for the presence of a chromosome segment introduced by appropriate crosses from a second strain. A strain is usually not regarded as congenic unless there have been at least eight crosses to the inbred strain" (Snell, 1978). Eaves et al. (2002) utilised microarray profiling of a NOD mouse, four NOD-derived diabetes-resistant congenic strains, and two non-diabetic control strains, to explore functional links between genotype and phenotype for T1D, using a novel method for differential expression analysis.

Here, the objective is to examine the correlation network of the gene expresssion profiles, link that information with available interactions and ontologies, and study the emergent patterns. The analysis is performed as explained in **Article II**. The resulting network is shown in Figure 9 of **Article II**, in which, some relevant entities in network are indicated with their names.

The largest upregulated cluster is related to lipid and glucose metabolism. Interestingly, the upregulated BRCA1 and BRCA2 genes are also placed within this cluster. BRCA genes are associated with breast cancer, but are known to be highly expressed in spleen and associated with immune response. How these genes specifically relate to Type 1 Diabetes is unclear and requires further study. Another upregulated small cluster of genes is found to be associated with beta-cell proliferation, which is a known response to increased rate of beta-cell apoptosis in Type 1 Diabetes.

4.1.5 Interaction neighbourhood in experimental context

In the context-based mining example given in a previous section (4.1.3), unit weights were used for all interactions to define the context. In general, as described in section 3.1.6, choosing proper weights to represent a context is a difficult task and most often an exploratory undertaking. Experimental data, such as gene expression or metabolomics experiments, can also be utilized to define a specific context. In such cases the distance measure relating biological entities in the molecular profile space may correspond to the measure of co-expression (such as correlation coefficient) between different entities.

To demonstrate the use of *MegNet* based mining using gene co-expression to define the context, we constructed an integrated network of interaction neighbourhood of mouse proteins PPAR-γ, PPAR-α, PGC1-α, GPR40, GPR41, and GPR43, together with co-expression network based on the gene expression data from spleens of various NOD related strains of mice (Eaves et al., 2002) (**Article III**). We then used a combination of the correlation coefficient values and unit weights for the edges to define the conceptual space, and visualised the similarities using CDA mapping (Fig. 3 of **Article III**).

Interestingly, several tumor suppressor genes such as BRCA1 associated with PPAR-γ, are found in this mapping. This finding deserves further attention. Only recently a link between a specific tumor suppressor (LKB1) and diabetes has been established, linking cancer and physiological control of metabolism (Shaw et al., 2005).

4.2 Dynamic topology of integrated networks

In this section, we present a study of dynamic topology changes. In order to study the dynamic changes in the topology of functional modules in an integrated network, we developed a method called Topological Enrichment Analysis of Functional Subnetworks (TEAFS; section 3.2). TEAFS performs topological analysis of *MegNet* networks using transcriptomics data. We applied TEAFS to study dynamic responses of Oxidative stress in yeast, and validated the results with comprehensive *in vivo* metabolomics analysis (**Article V**).

4.2.1 Oxidative stress

Oxidants are normally produced by aerobic metabolism, but are produced at elevated rates under pathophysiological conditions. Oxidative stress is defined as an imbalance between oxidants and antioxidants in favour of the oxidants, potentially leading to damage (Sies, 1997). Oxidative stress has been implicated in many human diseases such as atherosclerosis, diabetes, cancers, cardiovascular diseases, Parkinson's diseases, Alzheimer's disease, and aging (Sohal and Weindruch, 1996; Christen, 2000; Aruoma, 1998; Maritim et al., 2002; Miwa et al., 2008). Therefore, studying the phenomenon of oxidative stress is very important.

Saccharomyces cerevisiae or Baker's yeast shares remarkable similarities with higher eucaryotes and has served as a valuable model organism (Botstein et al., 1997) in facilitating understanding of numerous human diseases. It has been used as a model organism to study the effects of oxidative stress on aging (Gonidakis and Longo, 2008).

4.2.2 Transcriptomic response to Oxidative stress

Gasch et al. (2000) studied the responses, in terms of changes in transcript abundances over time, of *Saccharomyces cerevisiae* to a set of diverse environmental stresses. This set includes oxidative stress which was induced by growing cells to early-log phase and adding Hydrogen peroxide (H_2O_2) to a concentration of 0.30mM. Samples from this culture were collected at 10, 20, 30, 40, 50, 60, 80, 100, and 120 minutes and analysed with two-colour DNA microarrays (Schena et al., 1995). The resultant pattern of mRNA level responses to H_2O_2 treatment were characterised by the strong induction of genes that are involved in the detoxification of H_2O_2 and superoxides such as superoxide dismutases, glutathione peroxidases, and thiol-specific antioxidants, as well as genes involved in oxidative and reductive reactions within the cell including thioredoxin, thioredoxin reductases, glutaredoxin, and glutathione reductase (Gasch et al., 2000).

4.2.3 Topological response to Oxidative stress

The goal of the study presented in **Article V** is to gain an increased understanding of the dynamic response of the oxidative stress at the level of biological function. We integrated the transcriptomic information with interactome topology, formulating a method called Topological Enrichment Analysis of Functional Subnetworks (TEAFS), and showed that TEAFS analysis of topological changes derived from transcriptomics changes outperformed the traditional Gene Set Ernichment Analysis (GSEA) which is based on transcriptomic changes alone (Subramanian et al., 2005).

Metabolomics study

In order to assess the validity of the results of TEAFS analysis, we performed comprehensive study of metabolite concentrations during the course of oxidative stress on yeast. We conducted the metabolomics experiments with protocols for cell cultivation and H_2O_2 treatment identical to Gasch et al. (2000). Lipid profiling, primary metabolite and fatty acid profiling were done with Ultra Performance Liquid Chromatography coupled with Mass Spectrometry (UPLC/MS), High Performance Liquid Chromatography and Mass Spectrometry (HPLC/MS) and Gas Chromatography (GC) respectively (Goodacre et al., 2004; Orešič et al., 2006).

TEAFS analysis

An outline of TEAFS analysis is depicted in Figure 1 of **Article V** and a brief summary is provided here. First, we constructed a yeast genome-wide interaction network by integrating information pertaining to protein-protein interactions (DIP) (Xenarios et al., 2000), metabolic reactions (KEGG) (Kanehisa et al., 2006), genes encoding the proteins (EMBL, UniProt) (Stoesser et al., 1999; Bairoch et al., 2005), and transcriptional regulation (TRANSFAC) (Wingender et al., 2000; Matys et al., 2006) using *MegNet*. Next, this network was reconstructed corresponding to each time point in the oxidative stress experiment, by employing the gene detection criteria (*i.e.,* presence/absence criteria) similar to (Luscombe et al., 2004), (**Article V**). Next, functional modules were identified by making use of Gene Ontology Biological Process term annotations (Ashburner et al., 2000) available in the UniProt protein database (Bairoch et al., 2005). Finally, the test statistic associated with TEAFS (**Article V**) was computed for each module, significance of the statistic was computed using a permutation test (Moore and McCabe, 2005), and False Discovery Rate q-values were computed to account for multiple hypothesis testing (Benjamini and Hochberg, 1995; Shaffer, 1995). The results of TEAFS were compared to those of GSEA.

Results

TEAFS found changes in modules involved in environmental stress responses including oxidative stress response. These modules include regulation of cell cycle and check points, response to DNA damage stimulus (*i.e.,* repair mechanisms), cell wall organization, pentose phosphate shunt, biosynthesis of stress protectors (*i.e.,* glycogen and trehalose), signal transduction pathways, post-translational modifications, regulation of transcription and vacuolar acidification. In comparison, GSEA failed to identify many relevant changes.

Analysis of primary metabolites revealed consistent increase in the levels of trehalose-6-phosphate and decrease in that of pyruvate and mannose-6-phosphate during oxidative stress with respect to *S.*

cerevisiae under normal conditions. Analysis of fatty acids revealed consistent increase in the levels of palmitic acid (C16:0) during oxidative stress with respect to *S. cerevisiae* under normal conditions and increase in relative palmitate concentration over time. Analysis of lipids revealed increase in average ceramide concentrations over time. Lipid level changes, especially the ceramide and phospholipid levels, and the changes in functional modules—particularly lipid metabolism, phospholipid biosynthesis and ceramide biosynthesis—detected by TEAFS are in mutual accordance (**Article V**).

Fatty acid analysis results showed a significant increase in the levels of palmitic acid (16:0). Palmitate is a precursor of de novo ceramide biosynthesis which involves fatty acid elongation. The fatty acid elongase 3-ketoreductase (IFA38), which is encoded by the gene YBR159w, was identified as a hub protein in our integrated reference network which was absent at all time points under oxidative stress. The YBR159w mutant shows characteristic accumulation of ceramides and related reactive sphingolipids similar to other mutants with defects in fatty acid elongation (Han et al., 2002). This may indicate that accumulation of palmitate, a substrate to elongase system involving IFA38, and subsequent accumulation of ceramides are in part consequences of IFA38 response to oxidative stress.

Chapter 5

Summary and Conclusions

Post-genomic molecular biology is a data-rich field of research, and a variety of such data is publicly available. The goals of this thesis are: developing methods to make use of the available data in order to produce knowledge and to aid mining of newly generated data. Seeking to provide data mining for systems biology, the research has been built up on integration of heterogenous types of data. This thesis has made contributions in the areas of data integration, visualisation, mining, and study of network dynamics, which are important elements of systems biology (Figure 1.1).

This thesis has presented an integrated database which is a consolidation of a number of heterogeneous biological databases, and a software system called *MegNet* that enables retrieval and visualization of biological relationships across the data sources. Data traversals, an approach to linking interactions of heterogeneous types based on the identity of proteins, form the basis for the construction of integrated networks of interactions retrieved from multiple databases, and the *MegNet* software system allows visualisation of the networks (**Article I**). Context-sensitive mining of the data is facilitated by representing the data as conceptual spaces (**Article IV**), and visualising the similarities using dimension reduction (projection) techniques. As demonstrated by the results presented in section 4.1, **Article I**, **Article II**, and **Article III**, *MegNet* based data mining approach may facilitate discovery of novel or unexpected relationships, formulation of new hypotheses, data annotation, interpretation of new experimental data, and construction and validation of new network-based models of biological systems. High-throughput experimental techniques of post-genomic era are poor at quality and reproducibility. Therefore, integration of multiple types of data is desirable for bringing more confidence into analysis. Gene function prediction using *MegNet* has the potential to offer higher confidence as well as context-sensitivity.

The study of dynamic topological response of *Saccharomyces cerevisiae* to oxidative stress **Article V**, was based on the integration of transcriptomics and interactomics to predict altered biological processes. Comprehensive metabolomics was used to validate the integrative analysis. In this study, by performing an integrative analysis of transcriptomics and interactomics data, we have showed that the connectivity of the *Saccharomyces cerevisiae* cellular network is being dynamically modulated in response to oxidative stress, leading to progressive accumulation of (lipo)toxic lipids such as ceramides (**Article V**, section 4.2). Our approach takes advantage of connectivity of functional modules in heterogeneous interactome network constructed by *MegNet* and shows that connectivity based approach is superior to traditional pathway analysis. The findings from this study establish the applicability of our network analysis strategy, and support the hypothesis that modelling of local network topology dynamics can be used as an effective tool to study the activity of biological modules.

Omics data is ever-expanding and this poses challenges to updation and curation of data in datawarehousing approaches for data integration such as ours. It is not possible to completely avoid these problems, but by taking standards-based approach to data integration, the overheads in *e.g.,* keeping the schemas up-to-date can be reduced to some extent. In this thesis, we have relied largely on XML-based biological data exchange standards such as PSI-MI (Hermjakob et al., 2004), SBML (Hucka et al., 2003), KGML (Kanehisa et al., 2004) and so on. However, the diversity of the data, and the fact that not all data sources adapt the standards forces us to create our own schemas and write ad-hoc parsers in many cases. As explained in **Articles I–II**, we adapt the source schemas directly if available or we try to keep the problems arising from frequent schema changes to the minimum by extracting only relevant parts of the data to our databases. We adapted a combination of multiple approaches in data integration. Although we imported all the databases to the local warehouse, the individual schemas were kept intact. We only created an additional semantic mapping called *maps* database to facilitate resolution of entities across databases, which often doesn't need to change even when a new data source is added. The integration of data across databases and sophisticated queries are handled by java programs in the middle tier.

The results of the data integration techniques such as data traversals presented in this thesis are all from yeast, human or mouse 4.1. But the technique of data traversals is applicable more broadly to any organism for which we have large scale protein annotations and interactomics data such as gene regulatory reactions, metabolic reactions and protein-protein interactions. As protein identifiers are the central to the data traversals, the data traversals are possible if the different interaction databases use consistent identifiers or if the identifiers used in the interaction databases can be mapped to one common type of identifiers. We used Uniprot accession numbers are the standard identifiers for proteins, mainly because Uniprot is the primary database for protein annotations.

It is well known that publicly available molecular biology data include many false positives and thus quality filtering of these data is essential. There are exhaustive manual curation efforts (Keshava Prasad et al., 2009) as well as sophisticated filtering approaches to address these problems (Wu et al., 2009). In this thesis, however, we have not considered these aspects. Using high quality source data would enhance the value of our approaches.

Future work

Some topics of future developments include: extending Topological Enrichment Analysis of Functional Subnetworks (TEAFS; sections 3.2, 4.2) to be applicable to many commonly used experimental designs, extending gene detection criteria for many commonly used experimental platforms, and more fundamentally, defining a variety of interesting topological enrichment hypotheses and defining new schemes for computing enrichment score, to address these new hypotheses.

In this work we considered a variety of heterogeneous data for integrative modelling. However, integrating larger variety of data will enhance our abilities in more accurately modelling the biological systems (Kim et al., 2010). For example, Bauer-Mehren et al. (2009) demonstrated that integration of Single Nucleotide Polymorphism (SNP) data enables us to study the impact of the functional effect of SNPs in the structure and dynamics of biological networks. Thus, integration of SNPs would clearly be a natural extension to our framework.

Bibliography

Achard, F., Vaysseix, G., and Barillot, E. (2001). XML, bioinformatics and data integration. *Bioinformatics*, 17(2):115–125.

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., George, R. A., Lewis, S. E., Richards, S., Ashburner, M., Henderson, S. N., Sutton, G. G., Wortman, J. R., Yandell, M. D., Zhang, Q., Chen, L. X., Brandon, R. C., Rogers, Y. H., Blazej, R. G., Champe, M., Pfeiffer, B. D., Wan, K. H., Doyle, C., Baxter, E. G., Helt, G., Nelson, C. R., Gabor, G. L., Abril, J. F., Agbayani, A., An, H. J., Andrews-Pfannkoch, C., Baldwin, D., Ballew, R. M., Basu, A., Baxendale, J., Bayraktaroglu, L., Beasley, E. M., Beeson, K. Y., Benos, P. V., Berman, B. P., Bhandari, D., Bolshakov, S., Borkova, D., Botchan, M. R., Bouck, J., Brokstein, P., Brottier, P., Burtis, K. C., Busam, D. A., Butler, H., Cadieu, E., Center, A., Chandra, I., Cherry, J. M., Cawley, S., Dahlke, C., Davenport, L. B., Davies, P., de Pablos, B., Delcher, A., Deng, Z., Mays, A. D., Dew, I., Dietz, S. M., Dodson, K., Doup, L. E., Downes, M., Dugan-Rocha, S., Dunkov, B. C., Dunn, P., Durbin, K. J., Evangelista, C. C., Ferraz, C., Ferriera, S., Fleischmann, W., Fosler, C., Gabrielian, A. E., Garg, N. S., Gelbart, W. M., Glasser, K., Glodek, A., Gong, F., Gorrell, J. H., Gu, Z., Guan, P., Harris, M., Harris, N. L., Harvey, D., Heiman, T. J., Hernandez, J. R., Houck, J., Hostin, D., Houston, K. A., Howland, T. J., Wei, M. H., Ibegwam, C., Jalali, M., Kalush, F., Karpen, G. H., Ke, Z., Kennison, J. A., Ketchum, K. A., Kimmel, B. E., Kodira, C. D., Kraft, C., Kravitz, S., Kulp, D., Lai, Z., Lasko, P., Lei, Y., Levitsky, A. A., Li, J., Li, Z., Liang, Y., Lin, X., Liu, X., Mattei, B., McIntosh, T. C., McLeod, M. P., McPherson, D., Merkulov, G., Milshina, N. V., Mobarry, C., Morris, J., Moshrefi, A., Mount, S. M., Moy, M., Murphy, B., Murphy, L., Muzny, D. M., Nelson, D. L., Nelson, D. R., Nelson, K. A., Nixon, K., Nusskern, D. R., Pacleb, J. M., Palazzolo, M., Pittman, G. S., Pan, S., Pollard, J., Puri, V., Reese, M. G., Reinert, K., Remington, K., Saunders, R. D., Scheeler, F., Shen, H., Shue, B. C., Sidn-Kiamos, I., Simpson, M., Skupski, M. P., Smith, T., Spier, E., Spradling, A. C., Stapleton, M., Strong, R., Sun, E., Svirskas, R., Tector, C., Turner, R., Venter, E., Wang, A. H., Wang, X., Wang, Z. Y., Wassarman, D. A., Weinstock, G. M., Weissenbach, J., Williams, S. M., WoodageT, Worley, K. C., Wu, D., Yang, S., Yao, Q. A., Ye, J., Yeh, R. F., Zaveri, J. S., Zhan, M., Zhang, G., Zhao, Q., Zheng, L., Zheng, X. H., Zhong, F. N., Zhong, W., Zhou, X., Zhu, S., Zhu, X., Smith, H. O., Gibbs, R. A., Myers, E. W., Rubin, G. M., and Venter, J. C. (2000). The genome sequence of Drosophila melanogaster. *Science*, 287(5461):2185–2195.

Aebersold, R. and Mann, M. (2003). Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207.

Albert, R. and Barabási, A. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, 215(3):403–410.

Anderson, M. S. and Bluestone, J. A. (2005). The NOD mouse: a model of immune dysregulation. *Annu Rev Immunol*, 23:447–485.

Apweiler, R., Martin, M. J., O'Donovan, C., Magrane, M., Alam-Faruque, Y., Antunes, R., Barrell, D., Bely, B., Bingley, M., Binns, D., Bower, L., Browne, P., Chan, W. M., Dimmer, E., Eberhardt, R., Fedotov, A., Foulger, R., Garavelli, J., Huntley, R., Jacobsen, J., Kleen, M., Laiho, K., Leinonen, R., Legge, D., Lin, Q., Liu, W., Luo, J., Orchard, S., Patient, S., Poggioli, D., Pruess, M., Corbett, M., di Martino, G., Donnelly, M., van Rensburg, P., Bairoch, A., Bougueleret, L., Xenarios, I., Altairac, S., Auchincloss, A., Argoud-Puy, G., Axelsen, K., Baratin, D., Blatter, M.-C., Boeckmann, B., Bolleman, J., Bollondi, L., Boutet, E., Quintaje, S. B., Breuza, L., Bridge, A., deCastro, E., Ciapina, L., Coral, D., Coudert, E., Cusin, I., Delbard, G., Doche, M., Dornevil, D., Roggli, P. D., Duvaud, S., Estreicher, A., Famiglietti, L., Feuermann, M., Gehant, S., Farriol-Mathis, N., Ferro, S., Gasteiger, E., Gateau, A., Gerritsen, V., Gos, A., Gruaz-Gumowski, N., Hinz, U., Hulo, C., Hulo, N., James, J., Jimenez, S., Jungo, F., Kappler, T., Keller, G., Lachaize, C., Lane-Guermonprez, L., Langendijk-Genevaux, P., Lara, V., Lemercier, P., Lieberherr, D., de Oliveira Lima, T., Mangold, V., Martin, X., Masson, P., Moinat, M., Morgat, A., Mottaz, A., Paesano, S., Pedruzzi, I., Pilbout, S., Pillet, V., Poux, S., Pozzato, M., Redaschi, N., Rivoire, C., Roechert, B., Schneider, M., Sigrist, C., Sonesson, K., Staehli, S., Stanley, E., Stutz, A., Sundaram, S., Tognolli, M., Verbregue, L., Veuthey, A.-L., Yip, L., Zuletta, L., Wu, C., Arighi, C., Arminski, L., Barker, W., Chen, C., Chen, Y., Hu, Z.-Z., Huang, H., Mazumder, R., McGarvey, P., Natale, D. A., Nchoutmboube, J., Petrova, N., Subramanian, N., Suzek, B. E., Ugochukwu, U., Vasudevan, S., Vinayaka, C. R., Yeh, L. S., and Zhang, J. (2010). The universal protein resource (uniprot) in 2010. *Nucleic Acids Res*, 38(Database issue):D142–D148.

Aruoma, O. I. (1998). Free radicals, oxidative stress, and antioxidants in human health and disease. *J Am Oil Chem Soc*, 75(2):199–212.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. *Nat Genet*, 25(1):25–29.

Bader, G. D., Betel, D., and Hogue, C. W. V. (2003). BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res*, 31(1):248–250.

Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N., and Yeh, L.-S. L. (2005). The Universal Protein Resource (UniProt). *Nucleic Acids Res*, 33(Database issue):D154–D159.

Barabási, A.-L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5(2):101–113.

Barker, W. C., Garavelli, J. S., Haft, D. H., Hunt, L. T., Marzec, C. R., Orcutt, B. C., Srinivasarao, G. Y., Yeh, L. S., Ledley, R. S., Mewes, H. W., Pfeiffer, F., and Tsugita, A. (1998). The PIR-International Protein Sequence Database. *Nucleic Acids Res*, 26(1):27–32.

Barrett, T., Suzek, T. O., Troup, D. B., Wilhite, S. E., Ngau, W.-C., Ledoux, P., Rudnev, D., Lash, A. E., Fujibuchi, W., and Edgar, R. (2005). NCBI GEO: mining millions of expression profiles–database and tools. *Nucleic Acids Res*, 33(Database issue):D562–D566.

Bauer-Mehren, A., Furlong, L. I., Rautschka, M., and Sanz, F. (2009). From snps to pathways: integration of functional effect of sequence variations on models of cell signalling pathways. *BMC Bioinformatics*, 10 Suppl 8:S6.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, 57(1):289–300.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Rapp, B. A., and Wheeler, D. L. (2000). GenBank. *Nucleic Acids Res*, 28(1):15–18.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2010). Genbank. *Nucleic Acids Res*, 38(Database issue):D46–D51.

Berardini, T. Z., Li, D., Huala, E., Bridges, S., Burgess, S., McCarthy, F., Carbon, S., Lewis, S. E., Mungall, C. J., Abdulla, A., Wood, V., Feltrin, E., Valle, G., Chisholm, R. L., Fey, P., Gaudet, P., Kibbe, W., Basu, S., Bushmanova, Y., Eilbeck, K., Siegele, D. A., McIntosh, B., Renfro, D., Zweifel, A., Hu, J. C., Ashburner, M., Tweedie, S., Alam-Faruque, Y., Apweiler, R., Auchinchloss, A., Bairoch, A., Barrell, D., Binns, D., Blatter, M.-C., Bougueleret, L., Boutet, E., Breuza, L., Bridge, A., Browne, P., Chan, W. M., Coudert, E., Daugherty, L., Dimmer, E., Eberhardt, R., Estreicher, A., Famiglietti, L., Ferro-Rojas, S., Feuermann, M., Foulger, R., Gruaz-Gumowski, N., Hinz, U., Huntley, R., Jimenez, S., Jungo, F., Keller, G., Laiho, K., Legge, D., Lemercier, P., Lieberherr, D., Magrane, M., O'Donovan, C., Pedruzzi, I., Poux, S., Rivoire, C., Roechert, B., Sawford, T., Schneider, M., Stanley, E., Stutz, A., Sundaram, S., Tognolli, M., Xenarios, I., Harris, M. A., Deegan, J. I., Ireland, A., Lomax, J., Jaiswal, P., Chibucos, M., Giglio, M. G., Wortman, J., Hannick, L., Madupu, R., Botstein, D., Dolinski, K., Livstone, M. S., Oughtred, R., Blake, J. A., Bult, C., Diehl, A. D., Dolan, M., Drabkin, H., Eppig, J. T., Hill, D. P., Ni, L., Ringwald, M., Sitnikov, D., Collmer, C., Torto-Alalibo, T., Laulederkind, S., Shimoyama, M., Twigger, S., DEustachio, P., Matthews, L., Balakrishnan, R., Binkley, G., Cherry, J. M., Christie, K. R., Costanzo, M. C., Engel, S. R., Fisk, D. G., Hirschman, J. E., Hitz, B. C., Hong, E. L., Krieger, C. J., Miyasato, S. R., Nash, R. S., Park, J., Skrzypek, M. S., Weng, S., Wong, E. D., Aslett, M., Chan, J., Kishore, R., Sternberg, P., Auken, K. V., Khodiyar, V. K., Lovering, R. C., Talmud, P. J., Howe, D., and Westerfield, M. (2010). The gene ontology in 2010: extensions and refinements. *Nucleic Acids Res*, 38(Database issue):D331–D335.

Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B., and Shao, Y. (1997). The complete genome sequence of Escherichia coli K-12. *Science*, 277(5331):1453–1462.

Blobel, G. (1985). Gene gating: a hypothesis. *Proc Natl Acad Sci U S A*, 82(24):8527–8529.

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res*, 32(Database issue):D267–D270.

Bolstad, B. M., Collin, F., Simpson, K. M., Irizarry, R. A., and Speed, T. P. (2004). Experimental design and low-level analysis of microarray data. *Int Rev Neurobiol*, 60:25–58.

Botstein, D., Chervitz, S. A., and Cherry, J. M. (1997). Yeast as a model organism. *Science*, 277(5330):1259–1260.

Box, G. E. P., Hunter, W., and Hunter, S. (1969). *Statistics for experimenters*. John Wiley.

Breitkreutz, B.-J., Stark, C., Reguly, T., Boucher, L., Breitkreutz, A., Livstone, M., Oughtred, R., Lackner, D. H., Bähler, J., Wood, V., Dolinski, K., and Tyers, M. (2008). The biogrid interaction database: 2008 update. *Nucleic Acids Res*, 36(Database issue):D637–D640.

Butte, A. J. and Kohane, I. S. (2006). Creation and implications of a phenome-genome network. *Nat Biotechnol*, 24(1):55–62.

Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R., and Apweiler, R. (2004). The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res*, 32(Database issue):D262–D266.

Carreira-Perpiñan, M. A. (1997). A review of dimension reduction techniques. Technical Report CS-96-09, University of Sheffield, Sheffield, UK, CS-96-09, Sheffield, UK.

Casolari, J. M., Brown, C. R., Komili, S., West, J., Hieronymus, H., and Silver, P. A. (2004). Genome-wide localization of the nuclear transport machinery couples transcriptional status and nuclear organization. *Cell*, 117(4):427–439.

Ceol, A., Aryamontri, A. C., Licata, L., Peluso, D., Briganti, L., Perfetto, L., Castagnoli, L., and Cesareni, G. (2010). Mint, the molecular interaction database: 2009 update. *Nucleic Acids Res*, 38(Database issue):D532–D539.

Chatr-aryamontri, A., Ceol, A., Palazzi, L. M., Nardelli, G., Schneider, M. V., Castagnoli, L., and Cesareni, G. (2007). MINT: the Molecular INTeraction database. *Nucleic Acids Res*, 35(Database issue):D572–D574.

Chaudhri, A. B., Rashid, A., and Zicari, R. (2003). *XML data management: Native XML and XML-enabled database systems*. Addison-Wesley Professional.

Cherry, J. M., Adler, C., Ball, C., Chervitz, S. A., Dwight, S. S., Hester, E. T., Jia, Y., Juvik, G., Roe, T., Schroeder, M., Weng, S., and Botstein, D. (1998). Sgd: Saccharomyces genome database. *Nucleic Acids Res*, 26(1):73–79.

Christen, Y. (2000). Oxidative stress and Alzheimer disease. *Am J Clin Nutr*, 71(2):621S–629S.

Chuang, H.-Y., Lee, E., Liu, Y.-T., Lee, D., and Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Mol Syst Biol*, 3:140.

Codd, E. F. (1970). A relational model of data for large shared data banks. *Commun. ACM*, 13(6):377–387.

Curtis, R. K., Orešič, M., and Vidal-Puig, A. (2005). Pathways to the analysis of microarray data. *Trends Biotechnol*, 23(8):429–435.

Cusick, M. E., Yu, H., Smolyar, A., Venkatesan, K., Carvunis, A.-R., Simonis, N., Rual, J.-F., Borick, H., Braun, P., Dreze, M., Vandenhaute, J., Galli, M., Yazaki, J., Hill, D. E., Ecker, J. R., Roth, F. P., and Vidal, M. (2009). Literature-curated protein interaction datasets. *Nat Methods*, 6(1):39–46.

Demartines, P. and Hérault, J. (1997). Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets. *IEEE Transactions on Neural Networks*, 8(1):148–154.

Doubet, S., Bock, K., Smith, D., Darvill, A., and Albersheim, P. (1989). The complex carbohydrate structure database. *Trends Biochem Sci*, 14(12):475–477.

Drysdale, R. et al. (2008). Flybase : a database for the drosophila research community. *Methods Mol Biol*, 420:45–59.

Duarte, N. C., Becker, S. A., Jamshidi, N., Thiele, I., Mo, M. L., Vo, T. D., Srivas, R., and Palsson, B. Ø. (2007). Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc Natl Acad Sci U S A*, 104(6):1777–1782.

Duarte, N. C., Herrgård, M. J., and Palsson, B. Ø. (2004). Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model. *Genome Res*, 14(7):1298–1309.

Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18(1):71–103.

Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. (1998). *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge University Press.

Eaves, I. A., Wicker, L. S., Ghandour, G., Lyons, P. A., Peterson, L. B., Todd, J. A., and Glynne, R. J. (2002). Combining mouse congenic strains and microarray gene expression analyses to study a complex trait: the NOD model of type 1 diabetes. *Genome Res*, 12(2):232–243.

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*, 30(1):207–210.

Efron, B. and Tibshirani, R. (2006). On testing the significance of sets of genes. Technical report.

Ein-Dor, L., Kela, I., Getz, G., Givol, D., and Domany, E. (2005). Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, 21(2):171–178.

Engel, S. R., Balakrishnan, R., Binkley, G., Christie, K. R., Costanzo, M. C., Dwight, S. S., Fisk, D. G., Hirschman, J. E., Hitz, B. C., Hong, E. L., Krieger, C. J., Livstone, M. S., Miyasato, S. R., Nash, R., Oughtred, R., Park, J., Skrzypek, M. S., Weng, S., Wong, E. D., Dolinski, K., Botstein, D., and Cherry, J. M. (2010). Saccharomyces genome database provides mutant phenotype data. *Nucleic Acids Res*, 38(Database issue):D433–D436.

Fiehn, O. (2001). Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. *Comp Funct Genomics*, 2(3):155–168.

Flicek, P., Aken, B. L., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Fernandez-Banet, J., Gordon, L., Gräf, S., Haider, S., Hammond, M., Howe, K., Jenkinson, A., Johnson, N., Khri, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Koscielny, G., Kulesha, E., Lawson, D., Longden, I., Massingham, T., McLaren, W., Megy, K., Overduin, B., Pritchard, B., Rios, D., Ruffier, M., Schuster, M., Slater, G., Smedley, D., Spudich, G., Tang, Y. A., Trevanion, S., Vilella, A., Vogel, J., White, S., Wilder, S. P., Zadissa, A., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernndez-Suarez, X. M., Herrero, J., Hubbard, T. J. P., Parker, A., Proctor, G., Smith, J., and Searle, S. M. J. (2010). Ensembl's 10th year. *Nucleic Acids Res*, 38(Database issue):D557–D562.

Förster, J., Famili, I., Fu, P., Palsson, B. Ø., and Nielsen, J. (2003). Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. *Genome Res*, 13(2):244–253.

Garcia-Molina, H., Ullman, J. D., and Widom, J. (2002). *Database Systems: The Complete Book*. 0130319953. Prentice Hall PTR, Upper Saddle River, NJ, USA.

Gärdenfors, P. (2000). *Conceptual spaces: The geometry of thought*. The MIT Press.

Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D., and Brown, P. O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell*, 11(12):4241–4257.

Ge, H., Walhout, A. J. M., and Vidal, M. (2003). Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet*, 19(10):551–560.

Gelbart, W. M., Crosby, M., Matthews, B., Rindone, W. P., Chillemi, J., Twombly, S. R., Emmert, D., Ashburner, M., Drysdale, R. A., Whitfield, E., Millburn, G. H., de Grey, A., Kaufman, T., Matthews, K., Gilbert, D., Strelets, V., and Tolstoshev, C. (1997). Flybase: a drosophila database. the flybase consortium. *Nucleic Acids Res*, 25(1):63–66.

Gibney, M. J., Walsh, M., Brennan, L., Roche, H. M., German, B., and van Ommen, B. (2005). Metabolomics in human nutrition: opportunities and challenges. *Am J Clin Nutr*, 82(3):497–503.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H., and Oliver, S. G. (1996). Life with 6000 genes. *Science*, 274(5287):546, 563–546, 567.

Gonidakis, S. and Longo, V. D. (2008). *Oxidative Stress in Aging*, chapter Oxidative Stress and Aging in the Budding Yeast Saccharomyces cerevisiae, pages 67–79. Aging Medicine. Humana Press.

Goodacre, R., Vaidyanathan, S., Dunn, W. B., Harrigan, G. G., and Kell, D. B. (2004). Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends Biotechnol*, 22(5):245–252.

Goto, S., Okuno, Y., Hattori, M., Nishioka, T., and Kanehisa, M. (2002). LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res*, 30(1):402–404.

Han, G., Gable, K., Kohlwein, S. D., Beaudoin, F., Napier, J. A., and Dunn, T. M. (2002). The Saccharomyces cerevisiae YBR159w gene encodes the 3-ketoreductase of the microsomal fatty acid elongase. *J Biol Chem*, 277(38):35440–35449.

Hand, D., Mannila, H., and Smyth, P. (2001). *Principles of data mining*. The MIT Press.

Harris, T. W., Antoshechkin, I., Bieri, T., Blasiar, D., Chan, J., Chen, W. J., Cruz, N. D. L., Davis, P., Duesbury, M., Fang, R., Fernandes, J., Han, M., Kishore, R., Lee, R., Müller, H.-M., Nakamura, C., Ozersky, P., Petcherski, A., Rangarajan, A., Rogers, A., Schindelman, G., Schwarz, E. M., Tuli, M. A., Auken, K. V., Wang, D., Wang, X., Williams, G., Yook, K., Durbin, R., Stein, L. D., Spieth, J., and Sternberg, P. W. (2010). Wormbase: a comprehensive resource for nematode research. *Nucleic Acids Res*, 38(Database issue):D463–D467.

Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999). From molecular to modular cell biology. *Nature*, 402(6761 Suppl):C47–C52.

Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., Moore, S., Orchard, S., Sarkans, U., von Mering, C., Roechert, B., Poux, S., Jung, E., Mersch, H., Kersey, P., Lappe, M., Li, Y., Zeng, R., Rana, D., Nikolski, M., Husi, H., Brun, C., Shanker, K., Grant, S. G. N., Sander, C., Bork, P., Zhu, W., Pandey, A., Brazma, A., Jacq, B., Vidal, M., Sherman, D., Legrain, P., Cesareni, G., Xenarios, I., Eisenberg, D., Steipe, B., Hogue, C., and Apweiler, R. (2004). The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data. *Nat Biotechnol*, 22(2):177–183.

Herrgård, M. J., Swainston, N., Dobson, P., Dunn, W. B., Arga, K. Y., Arvas, M., Blüthgen, N., Borger, S., Costenoble, R., Heinemann, M., Hucka, M., Novère, N. L., Li, P., Liebermeister, W., Mo, M. L., Oliveira, A. P., Petranovic, D., Pettifer, S., Simeonidis, E., Smallbone, K., Spasi, I., Weichart, D., Brent, R., Broomhead, D. S., Westerhoff, H. V., Kirdar, B., Penttilä, M., Klipp, E., Palsson, B. Ø., Sauer, U., Oliver, S. G., Mendes, P., Nielsen, J., and Kell, D. B. (2008). A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nat Biotechnol*, 26(10):1155–1160.

Holden, M., Deng, S., Wojnowski, L., and Kulle, B. (2008). Gsea-snp: applying gene set enrichment analysis to snp data from genome-wide association studies. *Bioinformatics*, 24(23):2784–2785.

Hubbard, T. J. P., Aken, B. L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., Down, T., Dyer, S. C., Fitzgerald, S., Fernandez-Banet, J., Graf, S., Haider, S., Hammond, M., Herrero, J., Holland, R., Howe, K., Howe, K., Johnson, N., Kahari, A., Keefe, D., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Melsopp, C., Megy, K., Meidl, P., Ouverdin, B., Parker, A., Prlic, A., Rice, S., Rios, D., Schuster, M., Sealy, I., Severin, J., Slater, G., Smedley, D., Spudich, G., Trevanion, S., Vilella, A., Vogel, J., White, S., Wood, M., Cox, T., Curwen, V., Durbin, R., Fernandez-Suarez, X. M., Flicek, P., Kasprzyk, A., Proctor, G., Searle, S., Smith, J., Ureta-Vidal, A., and Birney, E. (2007). Ensembl 2007. *Nucleic Acids Res*, 35(Database issue):D610–D617.

Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., Arkin, A. P., Bornstein, B. J., Bray, D., Cornish-Bowden, A., Cuellar, A. A., Dronov, S., Gilles, E. D., Ginkel, M., Gor, V., Goryanin, I. I., Hedley, W. J., Hodgman, T. C., Hofmeyr, J.-H., Hunter, P. J., Juty, N. S., Kasberger, J. L., Kremling, A., Kummer, U., Novre, N. L., Loew, L. M., Lucio, D., Mendes, P., Minch, E., Mjolsness, E. D., Nakayama, Y., Nelson, M. R., Nielsen, P. F., Sakurada, T., Schaff, J. C., Shapiro, B. E., Shimizu, T. S., Spence, H. D., Stelling, J., Takahashi, K., Tomita, M., Wagner, J., Wang, J., and Forum, S. B. M. L. (2003). The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531.

Hulbert, E. M., Smink, L. J., Adlem, E. C., Allen, J. E., Burdick, D. B., Burren, O. S., Cassen, V. M., Cavnor, C. C., Dolman, G. E., Flamez, D., Friery, K. F., Healy, B. C., Killcoyne, S. A., Kutlu, B., Schuilenburg, H., Walker, N. M., Mychaleckyj, J., Eizirik, D. L., Wicker, L. S., Todd, J. A., and Goodman, N. (2007). T1DBase: integration and presentation of complex data for type 1 diabetes research. *Nucleic Acids Res*, 35(Database issue):D742–D746.

Ideker, T., Galitski, T., and Hood, L. (2001a). A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet*, 2:343–372.

Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R., and Hood, L. (2001b). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 292(5518):929–934.

Irizarry, R. A., Warren, D., Spencer, F., Kim, I. F., Biswal, S., Frank, B. C., Gabrielson, E., Garcia, J. G. N., Geoghegan, J., Germino, G., Griffin, C., Hilmer, S. C., Hoffman, E., Jedlicka, A. E., Kawasaki, E., Martnez-Murillo, F., Morsberger, L., Lee, H., Petersen, D., Quackenbush, J., Scott, A., Wilson, M., Yang, Y., Ye, S. Q., and Yu, W. (2005). Multiple-laboratory comparison of microarray platforms. *Nat Methods*, 2(5):345–350.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A*, 98(8):4569–4574.

Kaminuma, E., Mashima, J., Kodama, Y., Gojobori, T., Ogasawara, O., Okubo, K., Takagi, T., and Nakamura, Y. (2010). Ddbj launches a new archive database with analytical tools for next-generation sequence data. *Nucleic Acids Res*, 38(Database issue):D33–D38.

Kanehisa, M. and Bork, P. (2003). Bioinformatics in the post-sequence era. *Nat Genet*, 33 Suppl:305–310.

Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30.

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K. F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., and Hirakawa, M. (2006). From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*, 34(Database issue):D354–D357.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Res*, 32(Database issue):D277–D280.

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T. S., Hinrichs, A., Lu, Y. T., Roskin, K. M., Schwartz, M., Sugnet, C. W., Thomas, D. J., Weber, R. J., Haussler, D., Kent, W. J., and of California Santa Cruz, U. (2003). The ucsc genome browser database. *Nucleic Acids Res*, 31(1):51–54.

Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D. S., Sebastian, A., Rani, S., Ray, S., Harrys Kishore, C. J., Kanth, S., Ahmed, M., Kashyap, M. K., Mohmood, R., Ramachandra, Y. L., Krishna, V., Rahiman, B. A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R., and Pandey, A. (2009). Human Protein Reference Database–2009 update. *Nucl. Acids Res.*, 37(suppl 1):D767–772.

Kharchenko, P., Church, G. M., and Vitkup, D. (2005). Expression dynamics of a cellular metabolic network. *Mol Syst Biol*, 1:msb410023–E1:E6.

Killcoyne, S., Carter, G. W., Smith, J., and Boyle, J. (2009). Cytoscape: a community-based framework for network modeling. *Methods Mol Biol*, 563:219–239.

Kim, T. Y., Kim, H. U., and Lee, S. Y. (2010). Data integration and analysis of biological networks. *Curr Opin Biotechnol*.

Kitano, H. (2002). Systems biology: a brief overview. *Science*, 295(5560):1662–1664.

Krull, M., Pistor, S., Voss, N., Kel, A., Reuter, I., Kronenberg, D., Michael, H., Schwarzer, K., Potapov, A., Choi, C., Kel-Margoulis, O., and Wingender, E. (2006). TRANSPATH: an information resource for storing and visualizing signaling pathways and their pathological aberrations. *Nucleic Acids Res*, 34(Database issue):D546–D551.

Kulikova, T., Akhtar, R., Aldebert, P., Althorpe, N., Andersson, M., Baldwin, A., Bates, K., Bhattacharyya, S., Bower, L., Browne, P., Castro, M., Cochrane, G., Duggan, K., Eberhardt, R., Faruque, N., Hoad, G., Kanz, C., Lee, C., Leinonen, R., Lin, Q., Lombard, V., Lopez, R., Lorenc, D., McWilliam, H., Mukherjee, G., Nardone, F., Pastor, M. P. G., Plaister, S., Sobhany, S., Stoehr, P., Vaughan, R., Wu, D., Zhu, W., and Apweiler, R. (2007). Embl nucleotide sequence database in 2006. *Nucleic Acids Res*, 35(Database issue):D16–D20.

Lacroix, Z. and Critchlow, T. (2003). *Bioinformatics: managing scientific data*. Morgan Kaufmann San Francisco.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., Szustakowki, J., and Consortium, I. H. G. S. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Latchman, D. S. (1997). Transcription factors: an overview. *Int J Biochem Cell Biol*, 29(12):1305–1312.

Lee, J., Lendasse, A., and Verleysen, M. (2002). Curvilinear distance analysis versus isomap. In *Proceedings of European Symposium on Artificial Neural Networks (ESANN)*, pages 185–192, Bruges (Belgique).

Lee, J., Lendasse, A., and Verleysen, M. (2004). Nonlinear projection with curvilinear distances: Isomap versus curvilinear distance analysis. *Neurocomputing*, 57:49–76.

Lee, J. A., Lendasse, A., Donckers, N., and Verleysen, M. (2000). A robust non-linear projection method. In Verleysen, M., editor, *Proceedings European Symposium on Artificial Neural Networks (ESANN)*, pages 13–20, Bruges (Belgique).

Lemmens, I., Lievens, S., and Tavernier, J. (2010). Strategies towards high-quality binary protein interactome maps. *J Proteomics*.

Liu, M., Liberzon, A., Kong, S. W., Lai, W. R., Park, P. J., Kohane, I. S., and Kasif, S. (2007). Network-based analysis of affected biological processes in type 2 diabetes models. *PLoS Genet*, 3(6):e96.

Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol*, 14(13):1675–1680.

Lockhart, D. J. and Winzeler, E. A. (2000). Genomics, gene expression and DNA arrays. *Nature*, 405(6788):827–836.

Luscombe, N. M., Babu, M. M., Yu, H., Snyder, M., Teichmann, S. A., and Gerstein, M. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, 431(7006):308–312.

m Liu, W., Mei, R., Di, X., Ryder, T. B., Hubbell, E., Dee, S., Webster, T. A., Harrington, C. A., h Ho, M., Baid, J., and Smeekens, S. P. (2002). Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics*, 18(12):1593–1599.

Makino, S., Kunimoto, K., Muraoka, Y., Mizushima, Y., Katagiri, K., and Tochino, Y. (1980). Breeding of a non-obese, diabetic strain of mice. *Jikken Dobutsu*, 29(1):1–13.

Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O., and Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285(5428):751–753.

Maritim, A., Dene, B. A., Sanders, R. A., and Watkins, J. B. (2002). Effects of beta-carotene on oxidative stress in normal and diabetic rats. *J Biochem Mol Toxicol*, 16(4):203–208.

Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A. E., and Wingender, E. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*, 34(Database issue):D108–D110.

Michalickova, K., Bader, G. D., Dumontier, M., Lieu, H., Betel, D., Isserlin, R., and Hogue, C. W. V. (2002). SeqHound: biological sequence and structure database as a platform for bioinformatics research. *BMC Bioinformatics*, 3:32.

Michiels, S., Koscielny, S., and Hill, C. (2005). Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*, 365(9458):488–492.

Miwa, S., Beckman, K. B., and Muller, F. L., editors (2008). *Oxidative Stress in Aging*. Aging Medicine. Humana Press.

Montgomery, D. C. (1983). *Design and analysis of experiments*. Wiley.

Moore, D. S. and McCabe, G. P. (2005). *Introduction to the practice of statistics*, chapter Bootstrap Methods and Permutation Tests, pages 14.1–14.70. W. H. Freeman, 5 edition.

Mootha, V. K., Lindgren, C. M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrle, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D., and Groop, L. C. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*, 34(3):267–273.

Orešič, M., Clish, C. B., Davidov, E. J., Verheij, E., Vogels, J., Havekes, L. M., Neumann, E., Adourian, A., Naylor, S., van der Greef, J., and Plasterer, T. (2004). Phenotype characterisation using integrated gene transcript, protein and metabolite profiling. *Appl Bioinformatics*, 3(4):205–217.

Orešič, M., Hänninen, V. A., and Vidal-Puig, A. (2008). Lipidomics: a new window to biomedical frontiers. *Trends Biotechnol*, 26(12):647–652.

Orešič, M., Vidal-Puig, A., and Hänninen, V. (2006). Metabolomic approaches to phenotype characterization and applications to complex diseases. *Expert Rev Mol Diagn*, 6(4):575–585.

Papin, J. A. and Palsson, B. Ø. (2004). Topological analysis of mass-balanced signaling networks: a framework to obtain network properties including crosstalk. *J Theor Biol*, 227(2):283–297.

Patterson, S. D. and Aebersold, R. H. (2003). Proteomics: the first decade and beyond. *Nat Genet*, 33 Suppl:311–323.

Persico, M., Ceol, A., Gavrila, C., Hoffmann, R., Florio, A., and Cesareni, G. (2005). HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms. *BMC Bioinformatics*, 6 Suppl 4:S21.

Philippi, S. and Köhler, J. (2004). Using XML technology for the ontology-based semantic integration of life science databases. *IEEE Trans Inf Technol Biomed*, 8(2):154–160.

Quackenbush, J. (2002). Microarray data normalization and transformation. *Nat Genet*, 32 Suppl:496–501.

Raubal, M. (2004). Formalizing conceptual spaces. In *Formal ontology in information systems: proceedings of the third conference (FOIS-2004)*, page 153.

Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, B.-J., Hon, G. C., Myers, C. L., Parsons, A., Friesen, H., Oughtred, R., Tong, A., Stark, C., Ho, Y., Botstein, D., Andrews, B., Boone, C., Troyanskya, O. G., Ideker, T., Dolinski, K., Batada, N. N., and Tyers, M. (2006). Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae. *J Biol*, 5(4):11.

Rhead, B., Karolchik, D., Kuhn, R. M., Hinrichs, A. S., Zweig, A. S., Fujita, P. A., Diekhans, M., Smith, K. E., Rosenbloom, K. R., Raney, B. J., Pohl, A., Pheasant, M., Meyer, L. R., Learned, K., Hsu, F., Hillman-Jackson, J., Harte, R. A., Giardine, B., Dreszer, T. R., Clawson, H., Barber, G. P., Haussler, D., and Kent, W. J. (2010). The ucsc genome browser database: update 2010. *Nucleic Acids Res*, 38(Database issue):D613–D619.

Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A., and Chinnaiyan, A. M. (2004). ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia*, 6(1):1–6.

Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., and Séraphin, B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol*, 17(10):1030–1032.

Sammon, J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on computers*, 18(5):401–409.

Schacherer, F., Choi, C., Götze, U., Krull, M., Pistor, S., and Wingender, E. (2001). The TRANSPATH signal transduction database: a knowledge base on signal transduction networks. *Bioinformatics*, 17(11):1053–1057.

Schena, M., Heller, R. A., Theriault, T. P., Konrad, K., Lachenmeier, E., and Davis, R. W. (1998). Microarrays: biotechnology's discovery platform for functional genomics. *Trends Biotechnol*, 16(7):301–306.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470.

Seppänen-Laakso, T. and Orešič, M. (2009). How to study lipidomes. *J Mol Endocrinol*, 42(3):185–190.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annu Rev Psychol*, 46(1):561–584.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11):2498–2504.

Shaw, R. J., Lamia, K. A., Vasquez, D., Koo, S.-H., Bardeesy, N., Depinho, R. A., Montminy, M., and Cantley, L. C. (2005). The kinase LKB1 mediates glucose homeostasis in liver and therapeutic effects of metformin. *Science*, 310(5754):1642–1646.

Shoemaker, B. A. and Panchenko, A. R. (2007a). Deciphering protein-protein interactions. Part I. Experimental techniques and databases. *PLoS Comput Biol*, 3(3):e42.

Shoemaker, B. A. and Panchenko, A. R. (2007b). Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS Comput Biol*, 3(4):e43.

Sies, H. (1997). Oxidative stress: oxidants and antioxidants. *Exp Physiol*, 82(2):291–295.

Skusa, A., Rüegg, A., and Köhler, J. (2005). Extraction of biological interaction networks from scientific literature. *Brief Bioinform*, 6(3):263–276.

Snell, G. (1978). *Origins of inbred mice: proceedings of a workshop, Bethesda, Maryland, February 14-16, 1978*, chapter Congenic resistant strains of mice. Academic Press.

Sohal, R. S. and Weindruch, R. (1996). Oxidative stress, caloric restriction, and aging. *Science*, 273(5271):59–63.

Song, D. and Bruza, P. (2003). Towards context sensitive information inference. *Journal of the American Society for Information Science and Technology*, 54(4):321–334.

Speed, T. P., editor (2003). *Statistical analysis of gene expression microarray data.* CRC Press.

Spellman, P. T., Miller, M., Stewart, J., Troup, C., Sarkans, U., Chervitz, S., Bernhart, D., Sherlock, G., Ball, C., Lepage, M., Swiatek, M., Marks, W. L., Goncalves, J., Markel, S., Iordan, D., Shojatalab, M., Pizarro, A., White, J., Hubley, R., Deutsch, E., Senger, M., Aronow, B. J., Robinson, A., Bassett, D., Stoeckert, C. J., and Brazma, A. (2002). Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol*, 3(9):RESEARCH0046.

Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Res*, 34(Database issue):D535–D539.

Stein, L., Sternberg, P., Durbin, R., Thierry-Mieg, J., and Spieth, J. (2001). Wormbase: network access to the genome and biology of caenorhabditis elegans. *Nucleic Acids Res*, 29(1):82–86.

Stoesser, G., Tuli, M. A., Lopez, R., and Sterk, P. (1999). The EMBL Nucleotide Sequence Database. *Nucleic Acids Res*, 27(1):18–24.

Stuart, J. M., Segal, E., Koller, D., and Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–255.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545–15550.

Swanson, D. R. (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect Biol Med*, 30(1):7–18.

Tan, P. K., Downey, T. J., Spitznagel, E. L., Xu, P., Fu, D., Dimitrov, D. S., Lempicki, R. A., Raaka, B. M., and Cam, M. C. (2003). Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res*, 31(19):5676–5684.

Tateno, Y., Imanishi, T., Miyazaki, S., Fukami-Kobayashi, K., Saitou, N., Sugawara, H., and Gojobori, T. (2002). DNA Data Bank of Japan (DDBJ) for genome scale research in life science. *Nucleic Acids Res*, 30(1):27–30.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22(22):4673–4680.

Tian, L., Greenberg, S. A., Kong, S. W., Altschuler, J., Kohane, I. S., and Park, P. J. (2005). Discovering statistically significant pathways in expression profiling studies. *Proc Natl Acad Sci U S A*, 102(38):13544–13549.

Tipton, K. and Boyce, S. (2000). History of the enzyme nomenclature system. *Bioinformatics*, 16(1):34–40.

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., and Rothberg, J. M. (2000). A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature*, 403(6770):623–627.

Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995). Serial analysis of gene expression. *Science*, 270(5235):484–487.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Miklos, G. L. G., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Francesco, V. D., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guig, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. (2001). The sequence of the human genome. *Science*, 291(5507):1304–1351.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., and Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887):399–403.

Wang, K., Li, M., and Bucan, M. (2007). Pathway-based approaches for analysis of genomewide association studies. *The American Journal of Human Genetics*, 81(6):1278–1283.

Warren, P., Taylor, D., Martini, P. G. V., Jackson, J., and Bienkowska, J. (2007). PANP - a new method of gene detection on oligonucleotide expression arrays. In *Proc. 7th IEEE International Conference on Bioinformatics and Bioengineering BIBE 2007*, pages 108–115.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S. E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M. R., Brown, D. G., Brown, S. D., Bult, C., Burton, J., Butler, J., Campbell, R. D., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla, A. T., Church, D. M., Clamp, M., Clee, C., Collins, F. S., Cook, L. L., Copley, R. R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K. D., Deri, J., Dermitzakis, E. T., Dewey, C., Dickens, N. J., Diekhans, M., Dodge, S., Dubchak, I., Dunn, D. M., Eddy, S. R., Elnitski, L., Emes, R. D., Eswara, P., Eyras, E., Felsenfeld, A., Fewell, G. A., Flicek, P., Foley, K., Frankel, W. N., Fulton, L. A., Fulton, R. S., Furey, T. S., Gage, D., Gibbs, R. A., Glusman, G., Gnerre, S., Goldman, N., Goodstadt, L., Grafham, D., Graves, T. A., Green, E. D., Gregory, S., Guig, R., Guyer, M., Hardison, R. C., Haussler, D., Hayashizaki, Y., Hillier, L. W., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I., Jaffe, D. B., Johnson, L. S., Jones, M., Jones, T. A., Joy, A., Kamal, M., Karlsson, E. K., Karolchik, D., Kasprzyk, A., Kawai, J., Keibler, E., Kells, C., Kent, W. J., Kirby, A., Kolbe, D. L., Korf, I., Kucherlapati, R. S., Kulbokas, E. J., Kulp, D., Landers, T., Leger, J. P., Leonard, S., Letunic, I., Levine, R., Li, J., Li, M., Lloyd, C., Lucas, S., Ma, B., Maglott, D. R., Mardis, E. R., Matthews, L., Mauceli, E., Mayer, J. H., McCarthy, M., McCombie, W. R., McLaren, S., McLay, K., McPherson, J. D., Meldrim, J., Meredith, B., Mesirov, J. P., Miller, W., Miner, T. L., Mongin, E., Montgomery, K. T., Morgan, M., Mott, R., Mullikin, J. C., Muzny, D. M., Nash, W. E., Nelson, J. O., Nhan, M. N., Nicol, R., Ning, Z., Nusbaum, C., O'Connor, M. J., Okazaki, Y., Oliver, K., Overton-Larty, E., Pachter, L., Parra, G., Pepin, K. H., Peterson, J., Pevzner, P., Plumb, R., Pohl, C. S., Poliakov, A., Ponce, T. C., Ponting, C. P., Potter, S., Quail, M., Reymond, A., Roe, B. A., Roskin, K. M., Rubin, E. M., Rust, A. G., Santos, R., Sapojnikov, V., Schultz, B., Schultz, J., Schwartz, M. S., Schwartz, S., Scott, C., Seaman, S., Searle, S., Sharpe, T., Sheridan, A., Shownkeen, R., Sims, S., Singer, J. B., Slater, G., Smit, A., Smith, D. R., Spencer, B., Stabenau, A., Stange-Thomann, N., Sugnet, C., Suyama, M., Tesler, G., Thompson, J., Torrents, D., Trevaskis, E., Tromp, J., Ucla, C., Ureta-Vidal, A., Vinson, J. P., Niederhausern, A. C. V., Wade, C. M., Wall, M., Weber, R. J., Weiss, R. B., Wendl, M. C., West, A. P., Wetterstrand, K., Wheeler, R., Whelan, S., Wierzbowski, J., Willey, D., Williams, S., Wilson, R. K., Winter, E., Worley, K. C., Wyman, D., Yang, S., Yang, S.-P., Zdobnov, E. M., Zody, M. C., and Lander, E. S. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562.

Webb, E. C. (1992). *Enzyme nomenclature 1992: recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes*. Academic Press.

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Geer, L. Y., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Ostell, J., Miller, V., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R. L., Tatusova, T. A., Wagner, L., and Yaschenko, E. (2007). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 35(Database issue):D5–12.

Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I., Matys, V., Meinhardt, T., Prss, M., Reuter, I., and Schacherer, F. (2000). TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res*, 28(1):316–319.

Wu, J., Vallenius, T., Ovaska, K., Westermarck, J., Mäkelä, T. P., and Hautaniemi, S. (2009). Integrated network analysis platform for protein-protein interactions. *Nat Methods*, 6(1):75–77.

Xenarios, I., Rice, D. W., Salwínski, L., Baron, M. K., Marcotte, E. M., and Eisenberg, D. (2000). DIP: the database of interacting proteins. *Nucleic Acids Res*, 28(1):289–291.

Xenarios, I., Salwínski, L., Duan, X. J., Higney, P., Kim, S.-M., and Eisenberg, D. (2002). DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res*, 30(1):303–305.

Xiong, Y., Miyamoto, N., Shibata, K., Valasek, M. A., Motoike, T., Kedzierski, R. M., and Yanagisawa, M. (2004). Short-chain fatty acids stimulate leptin production in adipocytes through the G protein-coupled receptor GPR41. *Proc Natl Acad Sci U S A*, 101(4):1045–1050.

Yu, H., Braun, P., Yildirim, M. A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., Hao, T., Rual, J.-F., Dricot, A., Vazquez, A., Murray, R. R., Simon, C., Tardivo, L., Tam, S., Svrzikapa, N., Fan, C., de Smet, A.-S., Motyl, A., Hudson, M. E., Park, J., Xin, X., Cusick, M. E., Moore, T., Boone, C., Snyder, M., Roth, F. P., Barabási, A.-L., Tavernier, J., Hill, D. E., and Vidal, M. (2008). High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–110.

Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M., and Cesareni, G. (2002). MINT: a Molecular INTeraction database. *FEBS Lett*, 513(1):135–140.

Zhang, M., Yao, C., Guo, Z., Zou, J., Zhang, L., Xiao, H., Wang, D., Yang, D., Gong, X., Zhu, J., Li, Y., and Li, X. (2008). Apparently low reproducibility of true differential expression discoveries in microarray studies. *Bioinformatics*, 24(18):2057–2063.

Article V is not included in the PDF version.
Please order the printed version to get the complete publication
(http://www.vtt.fi/publications/index.jsp).

Article I

Data integration and visualization system for enabling conceptual biology

In: Bioinformatics.
Vol. 21, Suppl. 1, 2005, pp. i177–i185.
Published by permission of Oxford University Press.

BIOINFORMATICS

Vol. 21 Suppl. 1 2005, pages i177–i185
doi:10.1093/bioinformatics/bti1015

Data integration and visualization system for enabling conceptual biology

Peddinti V. Gopalacharyulu[1], Erno Lindfors[1,]
Catherine Bounsaythip[1], Teemu Kivioja[1], Laxman Yetukuri[1],
Jaakko Hollmén[2] and Matej Orešič[1,*]

[1]VTT Biotechnology, PO Box 1500, Espoo, FIN-02044 VTT, Finland and
[2]Helsinki University of Technology, Laboratory of Computer and Information Science,
PO Box 5400, Espoo, FIN-02015 HUT, Finland

Received on January 15, 2005; accepted on March 27, 2005

ABSTRACT

Motivation: Integration of heterogeneous data in life sciences is a growing and recognized challenge. The problem is not only to enable the study of such data within the context of a biological question but also more fundamentally, how to represent the available knowledge and make it accessible for mining.

Results: Our integration approach is based on the premise that relationships between biological entities can be represented as a complex network. The context dependency is achieved by a judicious use of distance measures on these networks. The biological entities and the distances between them are mapped for the purpose of visualization into the lower dimensional space using the Sammon's mapping. The system implementation is based on a multi-tier architecture using a native XML database and a software tool for querying and visualizing complex biological networks. The functionality of our system is demonstrated with two examples: (1) A multiple pathway retrieval, in which, given a pathway name, the system finds all the relationships related to the query by checking available metabolic pathway, transcriptional, signaling, protein–protein interaction and ontology annotation resources and (2) A protein neighborhood search, in which given a protein name, the system finds all its connected entities within a specified depth. These two examples show that our system is able to conceptually traverse different databases to produce testable hypotheses and lead towards answers to complex biological questions.

Contact: matej.oresic@vtt.fi

1 INTRODUCTION

Historically, the decomposition of biology into different disciplines was necessary to tackle the complexity of life science systems by 'reducing' the degree of complexity down to the most basic level. With the advent of 'omics' revolution and systems biology, such separation of biology is becoming artificial (Blagosklonny and Pardee, 2002). In order to utilize the diverse life science knowledge, one first needs to address several practical and fundamental challenges of data integration. For example, different domain-specific naming conventions and vocabularies have been utilized both at the low level, such as genes and proteins, and the more complex entities, such as biological concepts. In order to be able to integrate data, one should therefore enable traversing across such diverse sources of information in an automated way.

From the early days of bioinformatics, several approaches for biological data integration have been developed. Well-known approaches include rule-based links, such as SRS (Etzold and Argos, 1993; Etzold *et al*., 1996), federated middleware frameworks, such as Kleisli system (Davidson *et al*., 1997; Chung and Wong, 1999), as well as wrapper-based solution using query optimization, such as IBM Discovery Link (Hass *et al*., 2001). In parallel, progress has been made to organize biological knowledge in a conceptual way by developing ontologies and domain-specific vocabularies (Ashburner *et al*., 2000; Bard and Rhee, 2004; Bodenreider, 2004). With the emergence of XML and Semantic Web technologies, the ontology-based approach to life science data integration has become more ostensible. In this context, data integration comprises problems like homogenizing the data model with schema integration, combining multiple database queries and answers, transforming and integrating the latter to construct knowledge based on underlying knowledge representation.

However, the ontology-based approach alone cannot resolve the practical problem of evolving concepts in biology, and its best promise lies in specialized domains and environments where concepts and vocabularies can be well controlled (Searls, 2005; Oresic *et al*., 2005). Neither can the ontologies alone resolve the problem of context, i.e. what may appear closely related in one context, may be further apart or unrelated in another (Gärdenfors, 2000). In this paper, we present our approach to data integration and context-based mining of biological data, which is based on the premise that relationships between biological

[*]To whom correspondence should be addressed.

entities can be represented as a complex network, with nodes being either low level (e.g. genes, compounds) or more complex entities, such as concepts (cell localization, biological processes), and with edges being relationships between them, either physical interactions or more complex relationships.

The paper is organized as follows: in Section 2, we describe the practical implementation of our three-tier data integration system and the design of the Java-based tool we developed for querying the data and visualizing complex relationships. In Section 3, we demonstrate the utility of the system with two query examples: (1) an integrated pathway retrieval and (2) a protein neighborhood search. In Section 4, we discuss the design and performance of the system as well as its future developments.

2 SYSTEMS AND METHODS

2.1 System design

Our data integration and visualization system is composed of three layers in which the data constitutes the back-end layer (Fig. 1). Schema mappings, ontology definitions and conceptual learning implementations occupy the middle tier and the user interface constitutes the front-end layer. The middle tier also comprises sets of algorithms and modules that process and display results of the query. Most of our local data are represented in XML format. The data are stored using XML data management system Tamino XML server (Software AG) in a Redhat Linux Advanced Server v2.1 environment. The databases are queried using Tamino XQuery (Fiebig and Schöning, 2004) which is an implementation of XQuery language. The queries are enabled through the Tamino Java API. For storing more voluminous data, such as gene-expression data and in house produced mass spectrometry data, we use Oracle 10*g* database server (Oracle, Inc.).

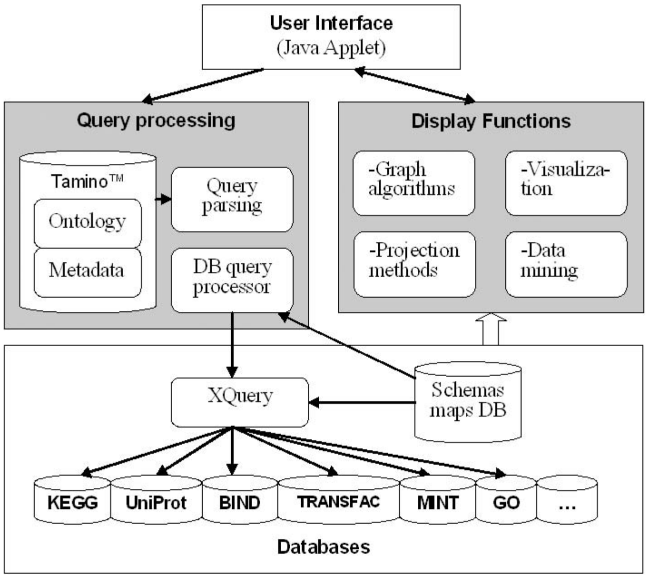


Fig. 1. Architecture of our bioinformatics data integration and visualization system.

2.2 Design of the network visualization tool

The megNet software is a Java-based tool which affords parallel retrieval across multiple databases, with results displayed as a network. Edge attributes contain information about types of relationships, possibly quantitative or semantic information (e.g. 'is located in' in case of linking a protein with a complex entity, such as cell organelle). The tool retrieves biological data from the Tamino databases using Tamino Java API and data from Oracle databases using JDBC. The user interface is implemented using Java Swing libraries, with the graphs created using Tom Sawyer Visualization Toolkit 6.0 (Tom Sawyer, Inc.). The basic layout of the user interface is divided into four parts (Fig. 2):

- query section,
- network display section,
- text area displaying information on currently selecting entity and
- distance mapping section, displaying the mapping of the distance matrix into 2D space.

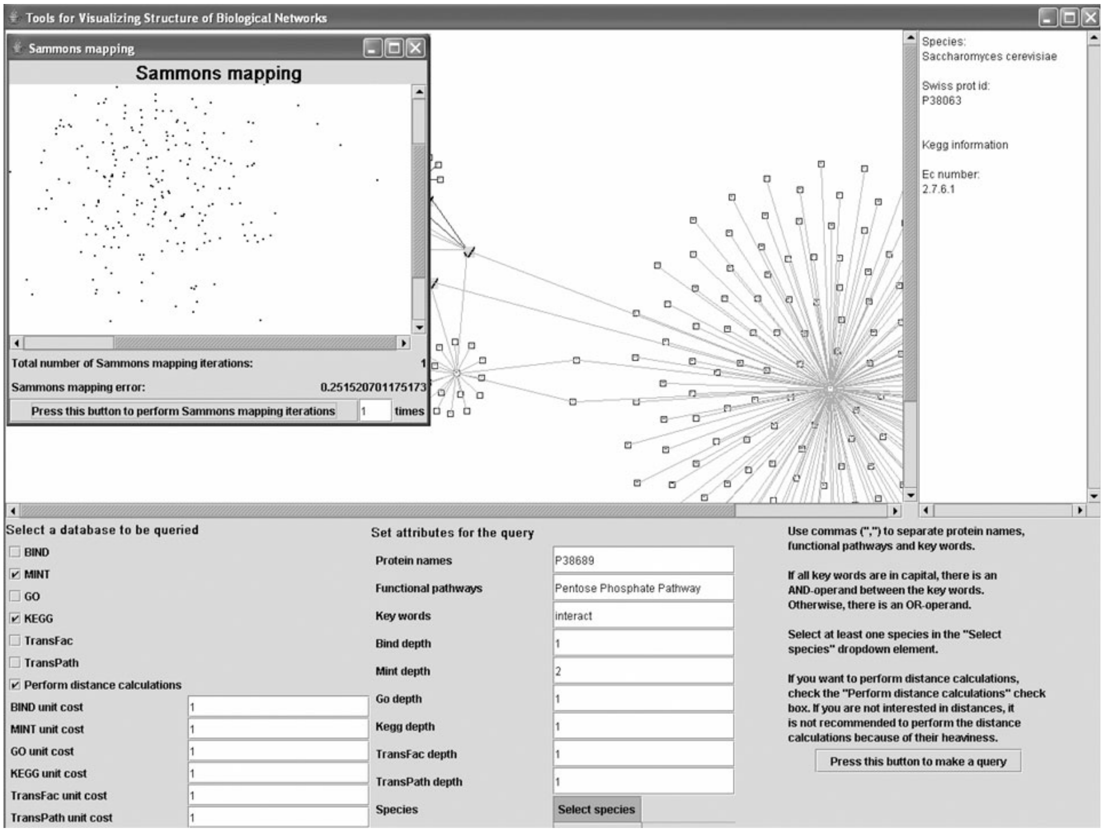
A mouse left click on a node or on an edge displays the biological information in the text area located on the right hand side. The information displayed in this text area contains the data retrieved from locally installed databases and links to external databases. The nodes can be selected to change options, such as set a new search depth for the neighbors. In the resultant graph, shape conventions are used to distinguish the type of entity underlying a node. Similarly, color codes are used to distinguish the type of relationship underlying an edge. Each node and edge shown can be checked for original source information. The resulting graph can be extracted and saved in the XML format.

2.3 Databases and data curation

Data from various public data sources were collected into our local database. Table 1 lists the data sources utilized in the examples of this paper.

In order to add a specific bioinformatics database into our system, it has to be passed first through a curation stage. A typical data curation flow is explained below in the form of a pseudoalgorithm:

(1) Decide on a data source to be set up and download the data typically using ftp. If the downloaded data are already in XML format go to step (3) otherwise go to (2).


Tools for Visualizing Structure of Biological Networks
Sammons mapping
Sammons mapping
Press this button to perform Sammons mapping iterations
times
Press this button to make a query
Select species

Species:
Saccharomyces cerevisiae

Swiss prot id:
P38063

Kegg information

Ec number:
2.7.6.1

Total number of Sammons mapping iterations: 1

Sammons mapping error: 0.251520701175173

Press this button to perform Sammons mapping iterations 1 times

Select a database to be queried

□ BIND
☑ MINT
□ GO
☑ KEGG
□ TransFac
□ TransPath
☑ Perform distance calculations

BIND unit cost 1
MINT unit cost 1
GO unit cost 1
KEGG unit cost 1
TransFac unit cost 1
TransPath unit cost 1

Set attributes for the query

Protein names P38689
Functional pathways Pentose Phosphate Pathway
Key words interact
Bind depth 1
Mint depth 2
Go depth 1
Kegg depth 1
TransFac depth 1
TransPath depth 1
Species Select species

Use commas (",") to separate protein names, functional pathways and key words.

If all key words are in capital, there is an AND-operand between the key words. Otherwise, there is an OR-operand.

Select at least one species in the "Select species" dropdown element.

If you want to perform distance calculations, check the "Perform distance calculations" check box. If you are not interested in distances, it is not recommended to perform the distance calculations because of their heaviness.

Fig. 2. Screenshot of the megNet network visualization tool. Node shapes represent their types (e.g. protein, gene), and edge colors represent types of relationships. The Sammon's mapping window displays the mapping based on specified distance metrics.

(2) Study the structure of the non-XML data and define XML schemas to capture the logical structure of the data. Go to step (4).

(3) If the document structures have been defined using DTD then convert the DTD to W3C Schema. If the XML schema is available from the source itself, if necessary, make changes to it to fit the requirements of the implementation (e.g. change the target namespace to Tamino namespace and define a prefix for the original target namespace).

(4) Define physical properties, such as indices and doctype for the logical schema to construct a Tamino Schema Definition document, i.e. TSD schema. If the previous step was (2) go to (5) or else go to (6).

(5) Develop parsers to convert the non-XML data into an XML format. A typical development phase is always followed by several test and feedback loops that involve an extensive use of XML data validation as well as human reading. Go to (7).

(6) Develop parsers to convert the distributed XML format to the required XML format.

(7) Load the resulting XML documents using mass-loading tool of the Tamino Server.

It must be noted that not every field in the source database is integrated. It is the task of the curator to capture its relevant subparts as well as to define appropriate semantics for the integrated database. Table 1 shows the XML Document Classes captured from databases used in this paper. In the course of implementing the above steps we make use of XMLSPY software (Altova, Inc.) and Tamino Schema Editor software (Software AG) for the construction and validation of logical and physical schemas, respectively. The development of parsers is usually implemented in Perl programming language and in some cases using Java.

2.4 Database traversals with schema maps

Resolving even simple biological relationships containing only a few biomolecular components often requires traversing multiple databases (Fig. 3). In order to enable such traversals within our system, we developed a database of schema maps (henceforth called maps database), which maps across different names used for the same entities across multiple databases. At the current state of development, the maps database

Table 1. Databases used in the present study

Database	Version or release date	XML document class	No. of entries
Uniprot/Swiss-Prot (Bairoch *et al*., 2005)	44.0	Uniprot	153 871
NCBI PubChem[a] (NCBI, 2004)	January 4, 2005	PC-substances	788 730
KEGG (Kanehisa *et al*., 2004)	August 2004	Pathways	11 380
LIGAND (Goto *et al*., 2002)		Gene	705 802
		Enzyme	4327
		Compound	11 116
		Glycan	10 302
TRANSFAC (Matys *et al*., 2003)	8.4	Gene	7796
		Factor	5919
		Site	14 782
TRANSPATH (Krull *et al*., 2003)	5.3	Network	72 769
Logical classes of data and entries:			
Pathway—333			
Gene—4989			
Molecule—20 164			
Reaction—23 065			
Annotation—24 218			
BIND (Bader *et al*., 2003)	August 27, 2004	BIND-submit	90 580
MINT (Zanzoni *et al*., 2002)	2.1	Entryset	18 951
IntAct (Hermjakob *et al*., 2004)	September 7, 2004	Entryset	37
Gene Ontology (Ashburner *et al*., 2000) assocdb XML version	January 4, 2004	GO	18 078

[a] NCBI PubChem (Accessed on January 10, 2005) http://pubchem.ncbi.nlm.nih.gov/

contains protein entities, indexed by UniProt identifiers. An example of such a map is shown in the XML code in Table 2. For creating such a map, we developed a Perl program to extract data from the Uniprot XML documents. We further extended this data with the GenInfo identifiers used in the BIND database (Bader *et al*., 2003) for each interacting protein. This data is obtained by applying the 'SeqHound-GetDefline' function of the SeqHound API (Michalickova *et al*., 2002). The HTTP method call for this 'SeqHound' function has been implemented using LWP module of the Perl programming language.

The database traversals can be achieved by applying simple join operations involving the maps database. Since the maps database records contain identifiers and names of an entity from all databases, it is ensured that the join operation between appropriate databases and rightly chosen entities would always return a non-empty result. The querying of a database independent of the names used in it can be achieved by writing queries to first search the maps database to find out the name/Id number of the entity in the original database and then search the original database with the correct name/Id number. Considerable challenge for any biological data integration is the often-changing structures of the data in the public databanks (Critchlow *et al*., 2000). We address this problem at the 'Logical schema construction level' of our data curation cycle by keeping our logical schemas to be as minimal as possible, yet useful enough to be able to observe the associations between all the data sources.

2.5 Similarity measures and graph projection

Property of similarity plays an essential role in human perception and formation of new concepts. The problem of evaluating similarity (or inversely, distance) between two entities or concepts appears more difficult when considering several 'quality dimensions' (Gärdenfors, 2000). In the domain of biology, the 'quality dimensions' could mean relationships of different types, i.e. chemical reactions, protein–protein interactions, gene sequence comparison or more complex relationships like protein localization, gene–phenotype association or compound properties.

Although distances within the molecular networks can be intuitively set to the length of the shortest path between the molecules, distance measure is less obvious for relationships, such as in ontologies. It was shown that Gene Ontology (GO) could be represented as a graph, and the distance measures in such a case were already studied (Lee *et al*., 2004). For the ontology trees, we assign a distance based on the closest common ancestor in the graph. When combining multiple relationships and corresponding distance measures, reasonable normalization of distance values has to be set in order to be able to compare across heterogeneous data sources. The distances between entities that do not have a direct relationship are then calculated as the

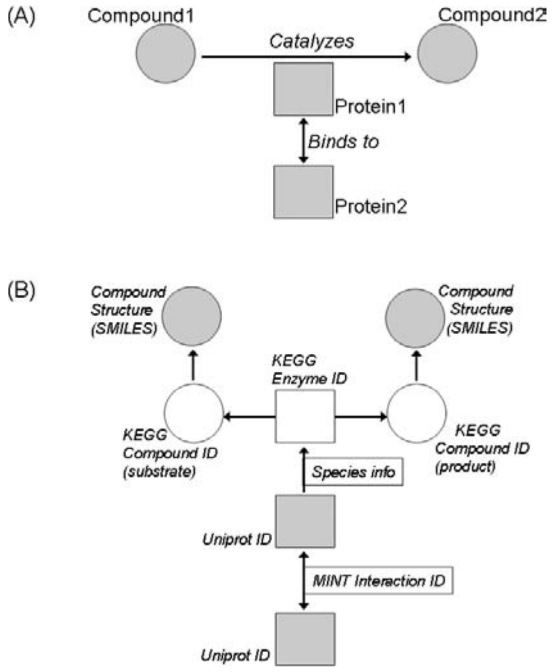


Fig. 3. (**A**) Schematic representation of relationships between two compounds and two proteins. (**B**) Same representation as hypothetically resolved via traversals across multiple databases.

lengths of the shortest paths with the distance-weighted edges (Fig. 4). The normalization of distances for each new data source is, in practice, handled by the bioinformaticians performing data curation. This assures that the system users do not need to know the specifics of the underlying data representation.

After distance normalization, it is ultimately up to the user to assign importance and therefore distance bias to any particular relationship type, by which context sensitivity can be achieved (Gärdenfors, 2000), as illustrated in Figure 4. When visualizing such complex data, we often need to project them into a lower dimensional space. In doing so it is important to preserve distances, i.e. two samples that are close to each other in the original space have to stay close when projected, or vice versa, two entities that are close to each other in the projected space must have come from the samples that were close to each other in the original space. It is the idea behind Sammon's mapping (Sammon Jr, 1969), which is implemented in our visualization tool. Visual configuration of entities is estimated with a gradient descent type of algorithm on a cost function based on the interpoint distances between the entities in the original space and the introduced discrepancies when applying the dimensionality-reducing mapping. In this way, the visual configuration approximates the original relationships in the complex networks. This kind of distance preservation is also used in the Kohonen's self-organizing maps (Kohonen, 2001) and multi-dimensional scaling (Torgerson, 1952).

Table 2. XML document from maps database for Uniprot protein entry AG35_VACCV, with links to indices from databases, such as EMBL, PIR, INTERPRO and Pfam

```
<?xml version="1.0" encoding="utf-8"?>
<protein created="1988-04-01" dataset="Swiss-Prot" ino:id="3426"
updated="2004-07-05">
  <primaryid>P07242</primaryid>
  <entry>AG35_VACCV</entry>
  <name>Envelope protein</name>
  <synonym>Protein H5</synonym>
  <synonym>Protein H6</synonym>
  <organism>
    <name>Vaccinia virus (strain WR)</name>
    <dbref id="10254" type="NCBI Taxonomy"/>
  </organism>
  <gene>
    <name>AG35</name>
    <synonym>H5R</synonym>
  <dbref id="M13209" type="EMBL">
    <property type="protein sequence ID"
    value="AAB59841.1"/>
  </dbref>
  <dbref id="M23648" type="EMBL">
  <property type="protein sequence ID"
    value="AAA47962.1"/>
  </dbref>
</gene>
<dblinks>
<dbref id="F24481" type="PIR">
  <property type="entry name" value="QQVZH6"/>
</dbref>
  <dbref id="IPR004966" type="InterPro">
  <property type="entry name" value="Pox_Ag35"/>
</dbref>
<dbref id="PF03286" type="Pfam">
  <property type="entry name" value="Pox_Ag35"/>
</dbref>
  <dbref id="138380" type="GenInfo"/>
</dblinks>
</protein>
```

3 EXAMPLES

3.1 Integrated pathway retrieval

Metabolic pathways and protein interaction networks have been studied extensively in the context of topology and modularity (Jeong *et al*., 2000, 2001). When attempting to model real biological phenomena, it is becoming clear that one needs to understand the cross-talk across different levels of biological organization, for example, between metabolic pathways and cell signaling (Papin and Palsson, 2004).

One of the primary motivations for the development of our bioinformatics system was the need to facilitate the study of available information in the context of biological questions.

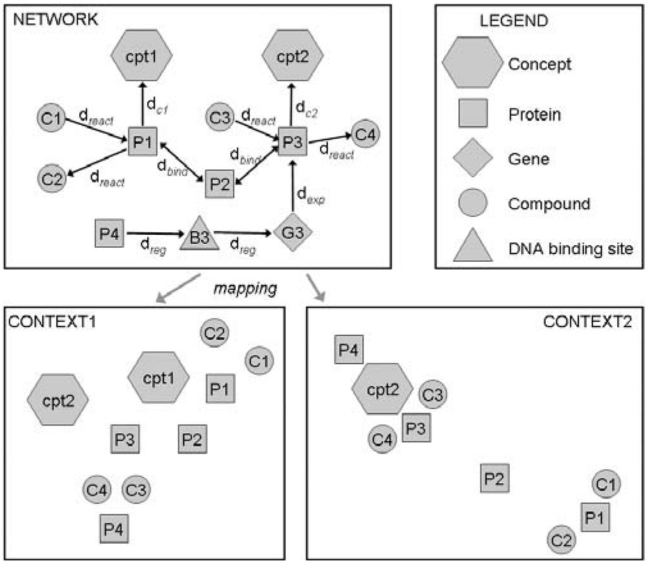


Fig. 4. Illustrative example of using graph projection in exploratory analysis of biological networks. In CONTEXT1 we are weighting all types of relationships similarly, so the nodes are clustered based on shortest path length between the edges. In CONTEXT2, we are interested only in concept cpt2, and assign lower distance value to nearest neighbors in metabolic pathways compared with other interactions.

One such application is the study of metabolic pathways, enriched with information about known molecular interactions at the level of protein–protein interactions, regulatory and signaling networks. As an example, we created the following query: ‘Glycolysis/Gluconeogenesis AND Pentose phosphate pathway AND TCA cycle IN *S.cerevisiae*’. The query was set up to first search the KEGG and retrieve the primary components of the pathways, i.e. enzymes and compounds. The database traversals were then used to search protein–protein interaction databases BIND and MINT for interactions of the enzymes with the nearest neighbor proteins (i.e. interaction search depth was set to 1). The resulting networks show surprisingly high level of connectivity across different stages of linear metabolic pathways via protein–protein interactions (Fig. 5). Specifically, in the zoomed-in region of Figure 5, we focus on two enzymes from the glycolysis pathway: phosphoglycerate kinase (PGK; EC 2.7.2.3) and acetate-CoA ligase (ACS; EC 6.2.1.1). ACS catalyzes formation of acetyl-CoA from acetate, which is a starting point in the TCA cycle, while PGK catalyzes acetylation of 3-phospho-D-glycerate, which is a part of the second phase of glycolysis. Both enzymes appear to aggregate with SRB2, based on the evidence from the yeast two-hybrid pooling approach (Ito *et al*., 2001). Notably, SRB2 is involved in transcriptional initiation (Thompson *et al*., 1993). This could mean that PGK and ACS, enzymes at two different stages of glycolysis, are coregulated. While the evidence from high-throughput yeast two-hybrid assays needs to be taken with caution due to possibly high number of false positive aggregation hits (Mrowka *et al*., 2001), our results do point toward a testable hypothesis for the future research.

3.2 Protein neighborhood search

Assignment of protein function is a non-trivial task owing to the fact that the same proteins may be involved in different biological processes, depending on the state of the biological system and protein localization (Camon *et al*., 2004). Therefore, protein function is context dependent.

The ‘protein neighborhood’, i.e. the entities of the network close to the protein, mode provide an insight about the protein function and its mode of action. The entities in our case can be molecules, genes or more complex concepts, and the proximity is measured by applying the distance measure. As an example, we searched the neighborhood of mannose-6-phosphate isomerase for *Saccharomyces cerevisiae* (PMI40; UniProt Id: P29952), which catalyzes the conversion between fructose 6-phosphate and mannose 6-phosphate and thus connects glycolysis with the cell wall synthesis in *S.cerevisiae* (Smith *et al*., 1992). The search involved concurrent retrieval of relationships for the following databases: UniProt, KEGG, BIND, MINT and GO Biological Process. For any nearest neighbor protein–protein association, such as protein–protein interaction or sharing the same GO class at the lowest level, the distance was set to 1. In the case of metabolic pathways, weight of each edge was set to 0.5 in the direction of possible reaction. The search depth was set to two nearest proteins if the first of the edges was a protein–protein interaction, and to the nearest protein otherwise. This included cases where the nearest protein was connected to the search protein via the compound in metabolic pathways or the lowest level GO term. Figure 6 shows the resulting graphs and Sammon’s mapping of the nearest protein neighbors of PMI40.

The zoomed-in window shows one region of potential interest, which includes protein–protein interactions between the PMI40 and NUP100 (UniProt Id: Q02629), a subunit of the nuclear pore complex, as well as between alpha-1,6-mannosyltransferase (MNN10; UniProt Id: P50108) and NUP100. According to GO (GO:0000032), both PMI40 and MNN10 are also involved in cell wall mannoprotein synthesis. While PMI40 is a ‘gate’ between cell wall synthesis and glycolysis, i.e. cell decision point between growth or energy production, MNN10 is a part of the protein complex in mannoprotein synthesis toward the end of the cell wall biosynthesis pathways. Examination of interaction entries (BIND Ids 137 955 and 137 823) suggests that NUP100 protein, which is a part of nuclear pore complex, binds to the PMI40 and MNN10 open reading frames (Casolari *et al*., 2004). This and other evidence by Casolari *et al*. provide support for the

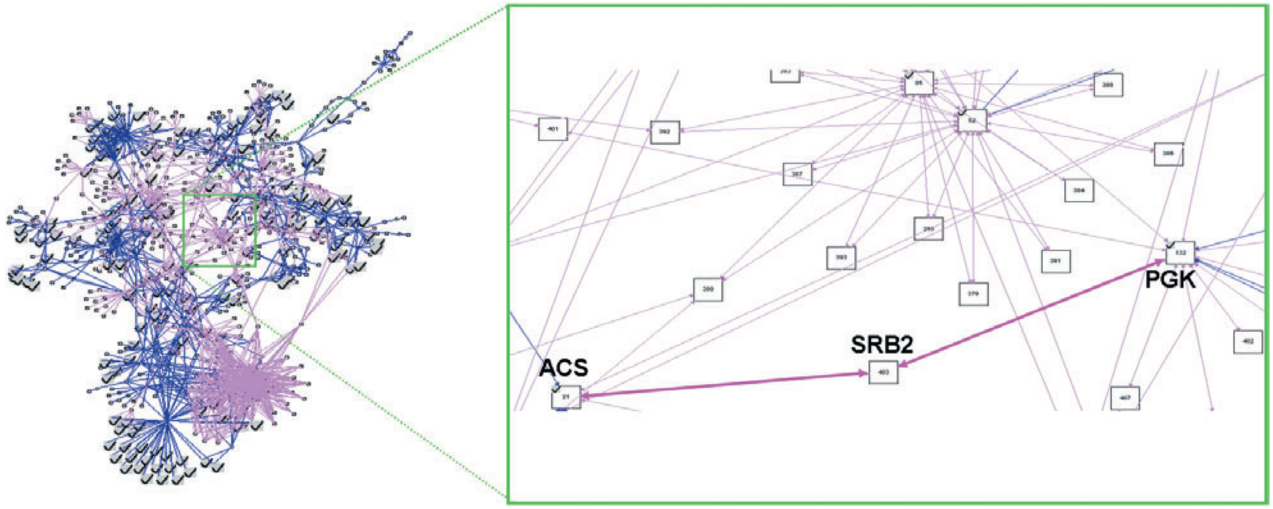


Fig. 5. Integrated pathway retrieval using megNet network visualization tool, with the query for 'Glycolysis/Gluconeogenesis AND Pentose phosphate pathway AND TCA cycle IN *S.cerevisiae*'. Metabolic pathways are shown with blue edges, protein–protein interactions with pink. Proteins are represented with squares, compounds with circles. Surprisingly, high level of connectivity via protein–protein interactions is found across different modules of the metabolism. The zoomed-in region shows a specific connection between Acetate-CoA ligase (ACS) and Phosphoglycerate kinase (PGK) via interactions with SRB2, which is known to be involved in transcriptional initiation. The interactions discussed are highlighted for clarity.

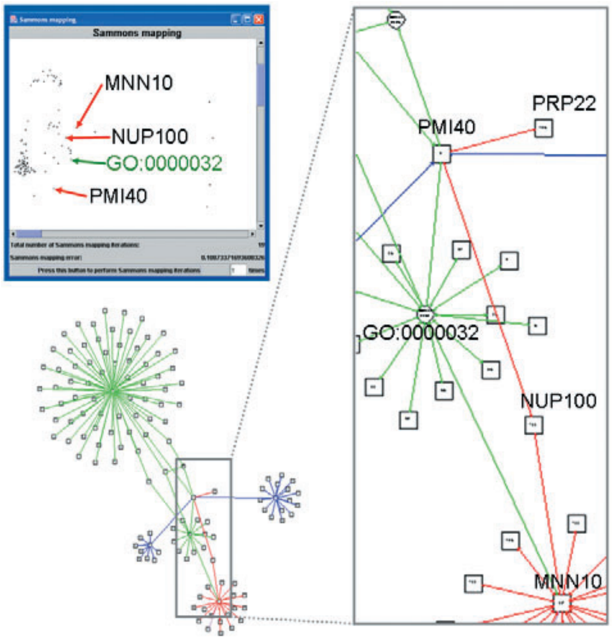


Fig. 6. Network neighborhood of mannose-6-phosphate isomerase (PMI40) in *S.cerevisiae*. Metabolic pathway relationships are shown in blue, protein–protein interactions in red, and GO associations in green. Both PMI40 and MNN10 are involved in cell wall mannoprotein synthesis (GO:0000032). NUP100 protein, which is part of the nuclear pore complex, appears to interact with the PMI40 and MNN10 genes.

'gene-gating' hypothesis, which suggests that the interaction of the nuclear pore complex with different genes might serve as a level of gene regulation (Blobel, 1985). It remains to be tested whether PMI40 and MNN10 are indeed coregulated in relation to cell decision-making between energy production versus growth.

4 DISCUSSION

Our integration approach is based on the premise that relationships between biological entities can be represented as a complex network. The information in such networks forms a basis for exploratory mining. Distances between different nodes in an integrated network play a central role in our framework. In order to calculate distances, one first needs to define distance measures across heterogeneous types of information. We are taking a pragmatic approach by letting the user define the distances as a part of the query. This is reasonable since the distance basically defines the context of the questions posed by the user and allows biasing the similarity toward particular types of relationships, or toward relationships in a specific context. Once the distance measure is specified, we can map the nodes of the graph into a lower dimensional space. As the mapping is approximate, there will be some distortion while doing the mapping. Therefore, in our opinion the exact form of distance measure is not a critical issue, so long as it underlines the relationships in the concept graph. In fact, selection of distance measure may reflect a subjective choice and as such will be subject to debate. It is ultimately the end result of mining that determines the utility of specific distance measure.

Presently, we are using Sammon's mapping for that purpose, which maps the graph non-linearly into lower dimensional space while preserving the internode distances across the network. One disadvantage of Sammon's mapping is that addition of the nodes requires new computation of the mapping on the complete network, and is therefore not well suited for interactive addition of new nodes. Other mappings, such as other types of multidimensional scaling methods (Torgerson, 1952) or self organizing maps (Kohonen, 2001), are also considered for future implementations. In particular, we will investigate the non-metric multidimensional scaling method (Cox and Cox, 2001), which is focused on preserving the order of similarities.

The two illustrative examples shown in the paper provide evidence for the usefulness of our approach. In the case of integrated pathway retrieval, we found large level of interconnectivity across different stages and modules of the metabolic pathways via protein–protein interactions, which raises questions about merit of studying the topology of metabolic networks outside the scope of other biological networks. Specifically, we found evidence of possible coregulation of enzymes at early and late stages of glycolysis pathway, which needs to be further investigated experimentally. In the case of protein neighborhood search, we were able to retrieve relationships and potential mechanisms that would not have been easily found through browsing databases separately. We believe our protein neighborhood search is a powerful tool for visual protein annotation in a context dependent manner.

Our approach is not limited to pathway databases and ontologies alone. We are currently extending the system in two directions. First, we aim at complementing the knowledge extracted from structured and semistructured data with the knowledge extracted from literature. Currently, we are implementing a text mining tool to retrieve from literature relationships between entities of interest, with primary focus on biomedical domain (Oresic *et al*., 2005). The discovered relationships will be, similarly as described in this paper, represented as a network. Second, genome information and experimental data such as metabolic profiles or gene-expression data can also be included. The distance measures in such cases are related to the level of association (e.g. correlation coefficient) or in the case of gene sequence comparison, to the alignment score. Combining molecular profile data with ontology information using database traversals has already been attempted (Oresic *et al*., 2004), but without the distance calculations.

We have presented an integrated database and software system that enables retrieval and visualization of biological relationships across heterogeneous data sources. We have demonstrated its merit on two practical examples: protein neighborhood search and integrated pathway retrieval. Owing to light-weight design of the system, it is relatively easy to incorporate new types of information and relationships. We believe our approach facilitates discovery of novel or unexpected relationships, formulation of new hypotheses, design of experiments, data annotation, interpretation of new experimental data, and construction and validation of new network-based models of biological systems.

ACKNOWLEDGEMENTS

M.O. was in part funded by Marie Curie International Reintegration Grant. M.O. and J.H. were in part funded by the Academy of Finland SYSBIO Programme.

REFERENCES

Ashburner,M., Ball,C., Blake,J., Botstein,D., Butler,H., Cherry,J., Davis,A., Dolinski,K., Dwight,S. and Eppig,J. (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Bader,G.D., Betel,D. and Hogue,C.W.V. (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.*, **31**, 248–250.

Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al*. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.

Bard,J.B.L. and Rhee,S.Y. (2004) Ontologies in biology: design, applications and future challenges. *Nat. Rev. Genet.*, **5**, 213–222.

Blagosklonny,M.V. and Pardee,A.B. (2002) Conceptual biology: unearthing the gems. *Nature*, **416**, 373.

Blobel,G. (1985) Gene gating: a hypothesis. *Proc. Natl Acad. Sci. USA*, **82**, 8527–8529.

Bodenreider,O. (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, **32**, D267–D270.

Camon,E., Magrane,M., Barrell,D., Lee,V., Dimmer,E., Maslen,J., Binns,D., Harte,N., Lopez,R. and Apweiler,R. (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res.*, **32**, D262–D266.

Casolari,J.M., Brown,C.R., Komili,S., West,J., Hieronymus,H. and Silver,P.A. (2004) Genome-wide localization of the nuclear transport machinery couples transcriptional status and nuclear organization. *Cell*, **117**, 427–439.

Chung,S.Y. and Wong,L. (1999) Kleisli: a new tool for data integration in biology. *Trends Biotechnol.*, **17**, 351–355.

Cox,T.F. and Cox,M.A.A. (2001) *Multidimensional Scaling*, Chapman and Hall/CRC, Boca Raton.

Critchlow,T., Fidelis,K., Ganesh,M., Musick,R. and Slezak,T. (2000) DataFoundry: information management for scientific data. *IEEE Trans. Inf. Technol. Biomed.*, **4**, 52–57.

Davidson,S.B., Overton,C.G., Tannen,V. and Wong,L. (1997) BioKleisli: a digital library for biomedical researchers. *Int. J. on Digital Libraries*, **1**, 36–53.

Etzold,T. and Argos,P. (1993) SRS—an indexing and retrieval tool for flat file data libraries. *CABIOS*, **9**, 49–57.

Etzold,T., Ulyanov,A. and Argos,P. (1996) SRS: information retrieval system for molecular biology data banks. *Methods enzymol.*, 114–128.

Fiebig,T. and Schöning,H. (2004) Software AG's Tamino XQuery Processor. *XIME-P 2004*, 19–24.

Goto,S., Okuno,Y., Hattori,M., Nishioka,T. and Kanehisa,M. (2002) LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res.*, **30**, 402–404.

Gärdenfors,P. (2000) *Conceptual Spaces: The Geometry of Thought*, MIT Press, Cambridge, MA.

Hass,L.M., Schwartz,P.M. and Kodali,P. (2001) DiscoveryLink: a system for integrated access to life science data sources. *IBM Systems Journal*, **40**, 489–511.

Hermjakob,H., Montecchi-Palazzi,L., Lewington,C., Mudali,S., Kerrien,S., Orchard,S., Vingron,M., Roechert,B., Roepstorff,P., Valencia,A. *et al*. (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res.*, **32**, D452–D455.

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

Jeong,H., Mason,S.P., Barabási,A.-L. and Oltvai,Z.N. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.

Jeong,H., Tombor,B., Albert, R., Oltvai,Z.N. and Barabási,A.-L. (2000) The large-scale organization of metabolic networks. *Nature*, **407**, 651–654.

Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.

Kohonen,T. (2001) *Self Organizing Maps*, Springer Verlag.

Krull,M., Voss,N., Choi,C., Pistor,S., Potapov,A. and Wingender,E. (2003) TRANSPATH: an integrated database on signal transduction and a tool for array analysis. *Nucleic Acids Res.*, **31**, 97–100.

Lee,S.G., Hur,J.U. and Kim,Y.S. (2004) A graph-theoretic modeling on GO space for biological interpretation of gene clusters. *Bioinformatics*, **20**, 381–388.

Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al*. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.

Michalickova,K., Bader,G., Dumontier,M., Lieu,H., Betel,D., Isserlin,R. and Hogue,C. (2002) SeqHound: biological sequence and structure database as a platform for bioinformatics research. *BMC Bioinformatics*, **3**, 32.

Mrowka,R., Patzak,A. and Herzel,H. (2001) Is there a bias in proteome research? *Genome Res.*, **11**, 1971–1973.

Oresic,M., Clish,C.B., Davidov,E.J., Verheij,E., Vogels,J.T.W.E., Havekes,L.M., Neumann,E., Adourian,A., Naylor,S., Greef,J.V.D. *et al*. (2004) Phenotype characterization using integrated gene transcript, protein and metabolite profiling. *Appl. Bioinformatics*, **3**, 205–217.

Oresic,M., Gopalacharyulu,P.V., Lindfors,E., Bounsaythip,C., Karanta,I., Hiirsalmi,M., Seitsonen,L. and Silvonen,P. (2005) Towards an integrative and context sensitive approach to *in silico* disease modelling. *ERCIM News*, 25–26.

Papin,J.A. and Palsson,B.O. (2004) Topological analysis of mass-balanced signaling networks: a framework to obtain network properties including crosstalk. *J. Theor. Biol.*, **227**, 283–297.

Sammon,J.W.Jr. (1969) A nonlinear mapping for data structure analysis. *IEEE Trans. Comp.*, **C-18**, 401–409.

Searls,D.B. (2005) Data integration: challenges for drug discovery. *Nat. Rev. Drug Disc.*, **4**, 45–48.

Smith,D., Proudfoot,A., Friedli,L., Klig,L., Paravicini,G. and Payton,M. (1992) PMI40, an intron-containing gene required for early steps in yeast mannosylation. *Mol. Cell. Biol.*, **12**, 2924–2930.

Thompson,C.M., Koleske,A.J., Chao,D.M. and Young,R.A. (1993) A multisubunit complex associated with the RNA polymerase II CTD and TATA-binding protein in yeast. *Cell*, **73**, 1361–1375.

Torgerson,W.S. (1952) Multidimensional scaling: I. theory and method. *Psychometrika*, **17**, 401–419.

Zanzoni,A., Montecchi-Palazzi,L., Quondam,M., Ausiello,G., Helmer-Citterich,M. and Cesareni,G. (2002) MINT: a molecular interaction database. *FEBS Lett.*, **513**, 135–140.

Article II

An integrative approach for biological data mining and visualisation

In: Int. J. Data Mining and Bioinformatics.
Vol. 2 (1), 2008, pp. 54–77.
Reprinted with permission from the publisher.

An integrative approach for biological data mining and visualisation

Peddinti V. Gopalacharyulu, Erno Lindfors, Jarkko Miettinen, Catherine K. Bounsaythip and Matej Orešič*

VTT Technical Research Centre of Finland,
P.O. Box 1500, Espoo, FIN-02044 VTT, Finland
E-mail: ext-gopal.peddinti@vtt.fi
E-mail: erno.lindfors@vtt.fi
E-mail: jarkko.miettinen@vtt.fi
E-mail: catherine.bounsaythip@vtt.fi
E-mail: matej.oresic@vtt.fi
*Corresponding author

Abstract: The emergence of systems biology necessitates development of platforms to organise and interpret plentitude of biological data. We present a system to integrate data across multiple bioinformatics databases and enable mining across various conceptual levels of biological information. The results are represented as complex networks. Context dependent mining of these networks is achieved by use of distances. Our approach is demonstrated with three applications: full metabolic network retrieval with network topology study, exploration of properties and relationships of a set of selected proteins, and combined visualisation and exploration of gene expression data with related pathways and ontologies.

Keywords: data mining; bioinformatics; complex networks; heterogeneous database integration; systems biology.

Reference to this paper should be made as follows: Gopalacharyulu, P.V., Lindfors, E., Miettinen, J., Bounsaythip, C.K. and Orešič, M. (2008) ‘An integrative approach for biological data mining and visualisation’, *Int. J. Data Mining and Bioinformatics*, Vol. 2, No. 1, pp.54–77.

Biographical notes: Peddinti V. Gopalacharyulu is a PhD student at the Helsinki University of Technology. He is pursuing his thesis work at VTT under the supervision of Matej Orešič. His research focuses on integration of heterogeneous biological data.

Erno Lindfors is embarking on his PhD studies at the Helsinki University of Technology. He is pursuing his thesis work at VTT under the supervision of Matej Orešič. His research focuses on visualisation of heterogeneous biological data.

Jarkko Miettinen is pursuing his Masters in a Bioinformatics Degree program at the Helsinki University of Technology.

Catherine K. Bounsaythip received her PhD in Automation and Computer Engineering from the University of Sciences and Technologies of Lille (France) for her work related to genetic algorithms. Her current research focuses on knowledge representation in biology.

Matej Orešič received his PhD in Biophysics from Cornell University, USA. His research interests include systems biology and metabolomics. He is a Group Leader of 'Quantitative Biology and Bioinformatics' at VTT.

1 Introduction

The *omics* revolution has empowered us with technologies to study the biological systems by measuring a large number of molecular components in parallel, therefore enabling the systems approach (Ideker et al., 2001; Kitano, 2002). The wealth of new information, combined with existing repositories of knowledge dispersed across numerous databases and literature, demand new solutions for management and integration of life science data. This has already been recognised in a variety of application domains relying on life science research. Knowledge management and data integration are recognised bottlenecks in drug discovery domain and current solutions are not yet capable of taking the full advantage of the information delivered by the modern *omics* technologies (Searls, 2005). More fundamentally, the ability to collect molecular information from biological systems in parallel is also challenging the ways we represent the biological systems and related knowledge, as well as the ways we design experiments to address specific biological questions.

Several approaches for biological data integration have been developed. Well-known examples include rule-based links such as SRS (Etzold and Argos, 1993; Etzold et al., 1996), federated middleware frameworks such as Kleisli system (Davidson et al., 1997; Chung and Wong, 1999), as well as wrapper-based solution using query optimisation such as IBM Discovery Link (Hass et al., 2001). In parallel, progress has been made to organise biological knowledge in a conceptual way by developing ontologies and domain-specific vocabularies (Ashburner et al., 2000; Bard and Rhee, 2004; Bodenreider, 2004). The emergence of XML and Semantic Web technologies has fostered the ontology-based approach to life science data integration. In this context, data integration comprises problems like homogenising the data model with schema integration, combining multiple database queries and answers, transforming and integrating the latter to construct knowledge based on underlying knowledge representation. However, the ontology-based approach alone cannot resolve the practical problem of evolving concepts in biology, and its best promise lies in specialised domains and environments where concepts and vocabularies can be well controlled. Neither can the ontologies alone resolve the problem of context, i.e., what may appear closely related in one context, may be further apart or unrelated in another (Gärdenfors, 2000).

Biological systems are characterised by the complexity of interactions of their internal parts and also with the external environment; integrating such information may result in a huge and heterogeneous network of biological entities. The visualisation of these networks poses many challenges (Herman et al., 2000). The problem is not only to display them, but also to represent them in a way that would enable easy interpretation of these huge networks. Our goal is to alleviate this problem by using context-based mining.

Biological network visualisation tools abound in many flavours, but few of them have met important requirements that enable real biological interpretation (Saraiya et al., 2005). Contextuality is one of those requirements. There are some tools that provide contextuality by attaching notes to visualised entities (Shannon et al., 2003; Dahlquist et al., 2002). However, this approach does not resolve the interpretation problem especially when the networks become complex. Therefore, the context-based mining is needed to eliminate some dimensions that are not contextually relevant.

Our approach to enable context-based mining is based on non-linear projection methods. Heterogeneous high-dimensional data are projected to a lower-dimensional space (two or three dimensions) in such a way that all similarity relationships are preserved as much as possible. This is quite challenging to implement in practice due to the heterogeneity of the entities and relationship types. The best compromise is to choose which kinds of relationships to visualise and what type of metrics to use in order to ensure the reliability and biological interpretability of the visualised data. Therefore, special attention should be put also on the data representation when integrating different types of information.

In this paper, we present a data integration and mining approach based on network representation models, which support an advanced visualisation system. As reported in our initial studies, the system has the capability to enable bioinformatics studies in a context dependent way (Gopalacharyulu et al., 2004, 2005). Section 2 introduces the general architecture of our database system, its implementation and methods. Section 3 describes our methods for network data representation and mining. Section 4 illustrates our approach on three different applications: metabolic network topology study, context-dependent protein annotation, and visualisation of Type 1 Diabetes gene expression dataset in the context of known pathways and ontologies. In the last section we discuss the current status of our research, persistent challenges, and future goals.

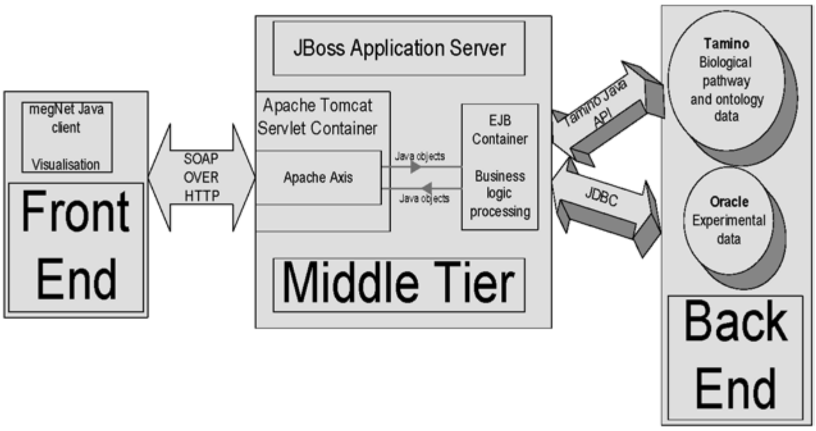
2 Integrated database system

2.1 Architectural design

The core architecture of our data integration and visualisation system, called *megNet*, is composed of three layers; back-end, middle tier and front-end (Figure 1). The data, schema maps, ontology definitions constitute the back-end layer. Most of our local data are represented in XML or RDF formats. The data is stored using XML data management system Tamino XML server (Software AG) in a Redhat Linux Advanced Server v3.0 environment. The databases are queried using Tamino X-Query which is based on XPath 1.0 specification. The queries are enabled through the Tamino Java API. For storing more voluminous data such as gene expression data and in house produced mass spectrometry data, we use Oracle 10g database server (Oracle, Inc.). The Oracle queries are performed using Oracle JDBC Thin drivers. The results obtained from queries to Tamino and Oracle are combined at the Java programming level in the middle tier.

The middle tier comprises the business logic of our system. Business logic events, such as graph *constructions*, *distance data projections*, *topology calculations* are implemented as stateless session beans. They are processed as web services. The session beans are the end points of the web services. They receive their request messages from the client for performing a business logic event. In the end of their life cycle they send the response to the client.

Figure 1 Three-tier architecture of the bioinformatics data integration and visualisation system. Back end tier consists of source biological data, schema mappings and ontologies. Middle tier is a suite of algorithms for business logic events (e.g., network constructions, data projections). Front end is a Java based user interface for visualisation the biological data and interacting with the user



The middle tier resides physically in a JBoss 4.04 Application Server (JBoss, Inc.). The business logic events are processed in the EJB Container of JBoss. The client and server communicate through SOAP messages. The SOAP messages are converted to Java objects by the middle tier after it has received a request message from the front-end client and Java objects are converted to SOAP messages before they are sent back as a response message. These conversions are implemented by using Apache Axis 1.4 (Apache Software Foundation). They are processed in Apache Tomcat 5.5 Servlet Container.

The front-end comprises the user interface for visualising and interacting with the end user. It is implemented in the Java environment.

2.2 Database curation

A system-wide life science data mining requires concurrent use of several databases, each of them likely having their own data schema, interface, address, and software tools. A database access tool is therefore needed that affords mining of several databases within one single interface. A fundamental step towards the integration of biological databases is to identify the 'atoms of information' and to develop solutions that resolve the naming conflicts as well as data structures. This is the task of a database 'curator'. For every database (either containing annotations or information about entity relationships) the database curator develops a data schema that enables mapping to other databases.

Data from various public and commercial data sources were set up in our database system. Table 1 lists those data sources which were utilised in the examples of this paper. A typical data curation flow is explained below in the form of a pseudo-algorithm:

1 Decide on a data source to be set up and download the data typically using ftp.
If the downloaded data is already XML format go to step (3) otherwise go to (2).

2 Study the structure of the non-XML data and define XML schemas to capture the logical structure of the data. Go to step (4). I

3 If the document structures have been defined using DTD, then convert the DTD to W3C Schema. If the XML schema is available from the source itself, if necessary, make changes to it to fit the requirements of the implementation (e.g., change the target name space to Tamino name-space and define a prefix for the original target namespace).

4 Define physical properties such as indices, doc-type etc. for the logical schema to construct a Tamino Schema Definition document, i.e., TSD schema. If the previous step was (2) go to (5) else go to (6).

5 Develop parsers to convert the non-XML data into an XML format. A typical development phase is always followed by several test and feed-back loops that involve an extensive use of XML data validation as well as human eye reading. Go to step (7).

6 Develop parsers to convert the distributed XML format to the required XML format.

7 Load the resulting XML documents using mass-loading tool of the Tamino Server.

Table 1 Databases incorporated into the system

Database	*Version or release date*	*No. of entries*
UniProt/Swiss-Prot (Bairoch et al., 2005)	44.0	153871
NCBI PubChem (http://pubchem.ncbi.nlm.nih.gov/)	January 4, 2005	–
Substance		788730
KEGG (Kanehisa et al., 2004)	August, 2004	–
Pathways		11380
LIGAND (Goto et al., 2002)		–
Genes		705802
Enzymes		4327
Compounds		11116
Glycans		10302
TRANSFAC (Matys et al., 2003)	June, 2005	–
Gene		7796
Factor		5919
Site		14782
TRANSPATH (Krull et al., 2003)	June, 2005	–
Pathway		333
Gene		4989
Molecule		20164
Reaction		23065
Annotation		24218
BIND (Bader et al., 2003)	August, 2004	90580
MINT (Zanzoni et al., 2002)	2.1	18951
IntAct (Hermjakob et al., 2004)	September, 2004	37
Gene Ontology (Gene Ontology Consortium, 2000) assocdb XML version	May, 2005	18078

As not every field in the original databases is integrated, it is the task of the curator to capture the relevant subparts of it as well as to define appropriate semantics for the integrated database. In the course of implementing the above steps we make use of XMLSPY software (Altova, Inc.) and Tamino Schema Editor software (Software AG) for the construction and validation of logical and physical schemas, respectively. The development of parsers is usually implemented in the Perl programming language and in some cases using Java.

2.3 Database traversals with schema maps

Even resolving simple biological relationships containing only a few biomolecular components often requires traversing multiple databases. In order to enable such traversals within our system, we developed a database of schema maps (henceforth called *maps* database), which maps across different names used for the same entities across multiple databases (Gopalacharyulu et al., 2005). For example, the maps database for protein entities is indexed by UniProt identifiers. For creating such a map, we developed a Perl program to extract data from the UniProt XML documents.

The database traversals can be achieved by applying simple join operations involving the maps database. Since the maps database records contain identifiers and names of an entity from all databases, it is ensured that the join operation between appropriate databases and rightly chosen entities would always return a non-empty result. The querying of a database independent of the names used in it can be achieved by writing queries to first search the maps database to find out the name/Id number of the entity in the original database and then search the original database with the correct name/Id number. Considerable challenge for any biological data integration is the often-changing structures of the data in the public databanks (Critchlow et al., 2000). We address this problem at the "Logical schema construction level" of our data curation cycle by keeping our logical schemas to be as minimal as possible, yet useful enough to be able to observe the associations between all the data sources.

3 Data visualisation and mining methods

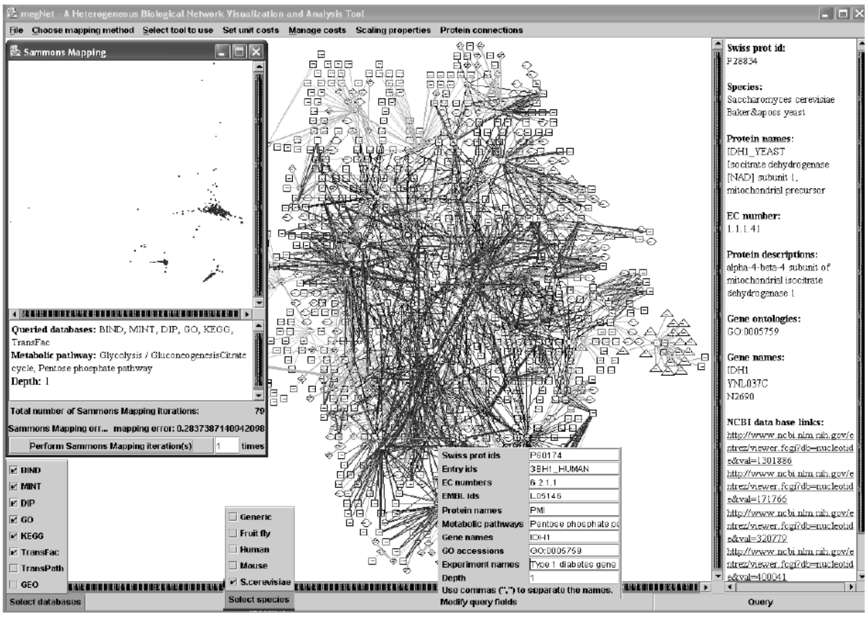
3.1 Network visualisation

In life sciences, everything is connected; even entities believed to be unrelated in some context might associate with each other in some other contexts. Thus, an integrated network of interacting entities of a biological system will necessarily contain many different types of entities and attributes arising from a number of disparate data sources, including literature databases.

The user interface of our system is capable of visualising these integrated networks in interactive manner (Figure 2). It constitutes the following sections:

- query parameters section
- network visualisation section
- display information section
- menu bar
- Non-Linear Mapping (NLM) window.

Figure 2 User interface of megNet, developed in Java



Queried databases: BIND, MINT, DIP, GO, KEGG, TransFac
Metabolic pathway: Glycolysis / GluconeogenesisCitrate cycle, Pentose phosphate pathway
Depth: 1

Total number of Sammons Mapping iterations: 79
Sammons Mapping err... mapping error: 0.2837387140942098
1 times

Swiss prot ids	P60174
Entry ids	3BH1_HUMAN
EC numbers	6.2.1.1
EMBL ids	L05146
Protein names	PMI
Metabolic pathways	Pentose phosphate pa
Gene names	IDH1
GO accessions	GO:0005759
Experiment names	Type 1 diabetes gene
Depth	1

Use commas (",") to separate the names.

Swiss prot id:
P28834

Species:
Saccharomyces cerevisiae
Baker&aposs yeast

Protein names:
IDH1_YEAST
Isocitrate dehydrogenase
[NAD] subunit 1,
mitochondrial precursor

EC number:
1.1.1.41

Protein descriptions:
alpha-4-beta-4 subunit of
mitochondrial isocitrate
dehydrogenase 1

Gene ontologies:
GO:0005759

Gene names:
IDH1
YNL037C
N2690

NCBI data base links:
http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&val=1301886
http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&val=171766
http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&val=320779
http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&val=400041

The ‘query parameters’ section consists of database, species, and query parameter menus. The database menu enables multiple selections from a list of all databases and the species menu enables multiple selections from a list of all species available in the system. The query parameter menu provides a collection of input boxes for entering a variety of parameters such as, protein names/ids, concept ids, metabolic pathway names, gene expression data set ids, initial depth of search etc. In addition, there is a button for launching the query.

The ‘network visualisation’ section is the place where the resulting network of a graph construction request is displayed. This interface provides options for interactively visualising or modifying the network. Typical examples of user interaction in this section include zooming in and out of the network, moving the network using pan tool, selecting a node to display its annotations in the display information section, selecting some parts of the network either to delete that part or to modify weights of the edges under selection etc.

The ‘display information’ section displays annotations of the selected node or edge. The information displayed reflects the annotations that exist in the databases. This section also provides hyperlinks to the source database of the entity under selection so as to enable the user to get more information on this entity.

The ‘menu bar’ enables interaction within our system in many ways. Typical example features enabled through its items include saving the network result or loading the network (in XML format), modifying weights of various types of interactions i.e., edge, projecting network into lower dimensional space and performing topological calculations on the networks.

The ‘NLM window’ displays the lower dimensional projection space. This interface also allows interactive features such as zooming in and out. Additionally, selecting a point in the projection space highlights the corresponding network node in the ‘network visualisation’ section. This enables viewing annotations of this entity in the ‘display information’ section.

When the user starts using the user interface, he can either load a previously saved network from XML document or he can construct a new network. In the former case he can open a file chooser from the upper menu for selecting the XML document. In the latter case he can assign query parameters to the network construction in the query parameter section that constitutes different menus on the bottom. In the database menu he can select from which databases he wants to retrieve entities and relationships. In the species menu, he selects in which species he wants to construct the network. In the query parameter menu, he can assign more parameters for the query. For example, he can type a protein name (e.g., PMI40) or identifier to visualise the neighbourhood of a certain protein. Or he can type a metabolic pathway name (e.g., Pentose phosphate pathway) to visualise all entities and interactions involved in a certain pathway or to investigate its neighbourhood of various types of interactions. When the user has assigned all query parameters, he can click on the ‘Query’ button to launch the query.

Once the network is constructed upon assigned query parameters or loaded from XML document, it is visualised on the middle part of the user interface (i.e., in Network visualisation section). The network is portrayed by using Tom Sawyer Visualisation 6.0 (Tom Sawyer Software, Oakland, CA, USA) symmetric layout algorithm. In the displayed network, shape conventions are used to distinguish the type of entity underlying a node. Similarly, colour codes are used to distinguish the type of the relationship underlying an edge. The user can make inferences from the network by zooming in and out. The user can save this network in XML format by opening a file chooser from the upper menu. A mouse left click on a node displays the biological information in the text area located on the right hand side. The information displayed in this text area contains the data retrieved from locally installed databases and links to external databases.

There are many ways to represent the data structure of a network (Bollobás, 1998). In our approach, a biological network is represented as a directed weighted graph where biological entities are nodes that are connected to each other through edges which are interactions or relationships between the entities. The shape of the nodes is coded differently depending on the type of an entity (e.g., squares stand for proteins, circles stand for compounds). The edges can be bidirectional or unidirectional, depending on the nature of the relationships. For example, in the case of protein-protein interaction network, we would relate the neighbouring proteins by searching all possible pathways among them, including their regulating genes. The generated nodes and edges then show the proteins and their interactions, respectively. In the case of metabolic network, we need to relate entities that are involved in each reaction. The substrates, products and enzymes are represented as nodes. As reactions can be either reversible or irreversible, unidirectional edges are used to distinguish the direction of an irreversible reaction and bidirectional edges are used to represent reversible reaction.

If the user wants to project the internal distances of the network into 2-dimensional space, she can assign appropriate bias by modifying the edge weights. After that she selects one of the available projection methods (Sammon’s NLM, Curvilinear Component Analysis (CCA), Curvilinear Distance Analysis (CDA)) from the upper menu (Each of these methods is described in detail in Section 3.2). After that the selected projection method is performed. As a result we obtain coordinates of the network nodes in the 2-dimensional projection space. These coordinates are displayed on a separate window that is opened after the projection method is finished. When the user clicks on a node on the two-dimensional projection window, the corresponding node on the network is highlighted and vice versa.

While distances within the molecular networks can be intuitively set to the length of the shortest path between the molecules, distance measure is less obvious for conceptual relationships such as in ontologies. One way to approach this is to consider an ontology as a graph and the distance measure is based on the shortest path to a common ancestor (Lee et al., 2004b). In the case of gene expression network which consists only of genes, the similarity measure is based on the gene expression profile distance between the genes (e.g., Euclidean or related).

The user can also perform topology calculations on the network and modify the network (e.g., removing some nodes according to their presence in an experimental condition). Our system uses a variety of methods for such studies. Below, we describe few that have been utilised in the examples of the paper.

3.2 Topology of a network

The molecular entities of the cell form a very complicated and dynamic interacting system. One of the major challenges of contemporary biology is to understand the structure of this complex web of interactions. The network structure and their dynamics is believed to have a significant effect on the structure and function of the cell (Barabasi and Oltvai, 2004).

The biological networks at the molecular level can be divided into different types of networks such as metabolic pathways, protein-protein interaction and regulatory networks. These networks are mutually interdependent and it has been demonstrated that they share some common network properties, e.g., the presence of single modularity networks (Barabasi and Oltvai, 2004; Han et al., 2004; Guimera and Amaral, 2005). However, the presence of the modularity in highly integrated biological networks is not self-evident as it lacks quantitative support (Ravasz and Barabási, 2003). There is thus a need for tools that afford the parallel study of multiple biological networks.

In order to study these topological properties we can formalise the network representation as a graph. Therefore, we apply mathematical methods used in graph theory.

Let us denote by $G = (X, U)$ a graph containing two sets where $X = \{x_1, x_2, \ldots, x_n, \ldots x_N\}_{|X|=N}$, the set of nodes and $U = \{u_1, u_2, \ldots, u_m, \ldots U_M\}_{|U|=M}$ the set of edges, where $u = [x_i, x_{i+1}]_{i=1\ldots N}$. A weighted graph is denoted by $G = (X, U, W)$ where $W: U \rightarrow \Re$.

The distances between the biological entities can be derived from the path lengths within a graph. A path μ of length q is a sequence of edges $U(\mu) = \{u_1, u_2, \ldots, u_q\}$. In a weighted graph the length of the path μ is obtained by summing up all weights of the edges of $U(\mu)$. In graphs, there are often many alternative paths between two nodes. Therefore, in practice one is mainly interested in the shortest path length between the selected nodes. We can obtain an average path length by calculating the shortest path between every pair of nodes of a graph and dividing the result by total number of nodes. This average value quantitatively characterises a graph by describing how close to each other its nodes are.

A graph can be characterised by its *degree distribution* $P_x(k)$ defining the probability that an arbitrary node x is connected to k neighbours. For metabolic networks, it was demonstrated that $P_x(k)$ decays as a power law $P_x(k) \approx k^{-\gamma}$ with $\gamma \cong 2.2$ in all organism (Jeong et al., 2000). This type of decay function characterises a *scale-free* network topology. This type of distribution is applicable only to a graph where all edges are bidirectional. For the case of networks containing some unidirectional edges, we would be interested in an *in-degree* distribution and *out-degree* distribution, which define the number of *in-coming* and *out-going* edges a node x has, respectively.

Another way to characterise a graph is to calculate its *clustering coefficient* $C_x(k)$ which is the density of connections in the neighbourhood of a node x (Dorogovtsev and Mendes, 2003). It is defined as the ratio between the total number n of the edges connected to its k nearest neighbours and the total number of all possible edges between all these nearest neighbours $C_x(k) = 2n/k(k - 1)$. A high clustering coefficient $C_x(k)$ would suggest a modular organisation.

It has been shown that most of complex networks (e.g., biological networks, world wide web, actor networks) are *scale free* networks with high *clustering coefficient* (Ravasz and Barabási, 2003). This means that there are few dominating hubs which lead to properties such as high tolerance to random failures. On the other hand, the network can collapse if one eliminates as few as 5–15% of its highly connected hubs. Recent studies showed that metabolic networks contain a *hierarchical modularity* (Kanehisa et al., 2004). This modularity combines two features into one network type. According to this modularity study, graph's *in-* and *out-degree* distributions follow power law $P_x(k) \approx k^{-\gamma}$, with a constant $\gamma \in \Re$, and the dependence of the clustering coefficient follows the power law $C_x(k) \approx k^{-\gamma}$ as well.

3.3 Network projections

The main purpose of data projection is to map a high dimensional data to a lower dimensional space in order to be able to visualise them in a context-based manner. The methods implemented in our system so far are the Sammon's NLM (Sammon, 1969), CCA (Demartines and Hérault, 1997) and CDA (Lee et al., 2004a).

All projection methods we used share common features:

Let d_{ij}^{*} denote distance, by some metric, between two points i and j in the original K-dimensional input space **A** and let d_{ij} denote the distance between points i and j in the L-dimensional (where $L < K$) output space **B**. In addition, every projection method we have used has an error function Err(.) which includes these two distances and some weight function which decides on how much smaller or larger distances we try to preserve.

All methods try to minimise an error function iteratively, either by steepest gradient descent (NLM) or stochastic gradient descent (CCA and CDA).

3.3.1 Sammon's Non-Linear Mapping (NLM)

Sammon's NLM (Sammon, 1969) error function is the following:

$$\text{Err} = \frac{1}{\sum_{i<j}^{K} d_{ij}^{*}} \sum_{i<j}^{K} \frac{(d_{ij}^{*} - d_{ij})^2}{d_{ij}^{*}}.$$

NLM algorithm tries to minimise Err by always descending towards the steepest gradient. It may thus end up in a local minimum and the convergence may be slow. Its time-complexity is of $O(n^2)$. Therefore it may be too slow for data with tens of thousands of points, especially when the original dimensionality K is large, and is not appropriate for interactive work.

3.3.2 Curvilinear Component Analysis (CCA)

CCA attempts to preserve local topology by favouring first short distances, and long distances afterwards. The error function is formalised as follows:

$$\text{Err} = \frac{1}{2}\sum_{i}\sum_{i\neq j}(d_{ij}^{*} - d_{ij})^{2} F(d_{ij}, \lambda(k))$$

where $F(d_{ij}, \lambda(k))$ is the weighting neighbourhood function that decreases with its arguments, thus favours local topology preservation. Computationally CCA is lighter than NLM because CCA reduces the computational cost of finding minima by using stochastic gradient descent and by optionally using vector quantisation to create centroids that approximate some groups of points in K-space. Without quantisation CCA's time-complexity is of $O(n^2)$ and with vector quantisation $O(n^*n')$ where n' is the number of centroids created in vector quantisation. Therefore, the time-complexity becomes $O(n^2)$ with inefficient vector quantisation.

3.3.3 Curvilinear Distance Analysis (CDA)

Instead of calculating Euclidean distances between points of an object, CDA calculates curvilinear distances, denoted by δ_{ij}, between points of a structure by creating a graph out of centroids. After that it calculates the shortest path between two prototypes of the codebook after quantisation and linking of the prototypes. The curvilinear distances are used instead of Euclidean distances. The error function becomes then:

$$\text{Err} = \frac{1}{2}\sum_{i}\sum_{i\neq j}(\delta_{ij}^{*} - \delta_{ij})^{2} F(d_{ij}, \lambda(k)).$$

CDA's time-complexity is of $O(n'e + n'^2\ln(n'))$, where e is number of edges created between centroids, n' number of centroids and n number of data-points. This follows from the complexity of Dijkstra's (1959) shortest path algorithm that is used for every centroid. That becomes $O(n.e + n^2\ln(n))$ with inefficient vector quantisation.

In the worst case the runtimes of CDA may seem to be very long compared to that of CCA or NLM. However, in practice its runtime is near that of CCA which is much shorter than that of NLM. The use of curvilinear distance measure provides much better results than CCA when K-space has complex features. In the following section, we will apply CDA projection method to visualise the metabolic network in a context-based manner.

4 Applications

4.1 Network retrieval and topology study

The topological properties of biological networks have been an intense topic of computational biology research (Jeong et al., 2000, 2001; Arita, 2004; Barabasi and Oltvai, 2004). A practical step necessary to retrieve specific networks involved in such studies requires development of parsers to retrieve those networks from appropriate databases. Since it is becoming clear the topology of biological network may also need to be viewed in the context of systems dynamics (Luscombe et al., 2004), the future research in this domain would benefit from ability to retrieve biological networks corresponding to different biological states easily from the life science databases and experimental data.

A simple example of a network retrieved from our database is presented in Figure 3, showing a result from a query for the complete metabolic network from KEGG (Kanehisa et al., 2004) for *S. cerevisiae* species. This network can then be investigated for local structures, links to other networks and biological entities, as well as for the global studies such as analyses of network scaling properties. Figure 4 shows the calculated degree distribution of the yeast metabolic network retrieved from KEGG, with the nodes being the enzymes and the edges connections between the enzymes via metabolites as substrates or products. Figure 5 shows the calculated degree distribution as a function of node degree for the same network. It appears that neither of these distributions follows the power law ideally, which is in contrast with previous findings stating that the hierarchical modularity is present in metabolic networks (Jeong et al., 2000). We can see from Figure 3 that there is one large metabolic island which contains most nodes of the graph. The presence of several small islands may be explained by the lack of the connectivity data in KEGG. These islands affect the total distributions.

Figure 3 Result of a retrieval of complete yeast metabolic network from megNet using a simple query for KEGG and *S. cerevisiae*

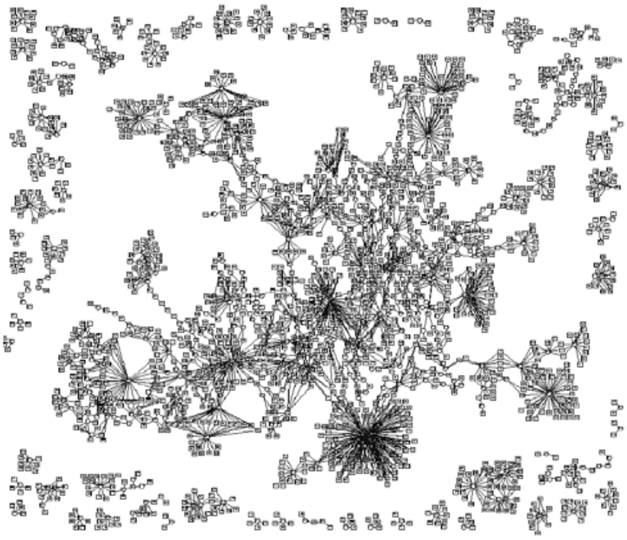


Figure 4 Degree distribution of the yeast metabolic network shown in Figure 3. It appears that the degree distribution does not follow the power law which means that there is no hierarchical modularity in this metabolic network

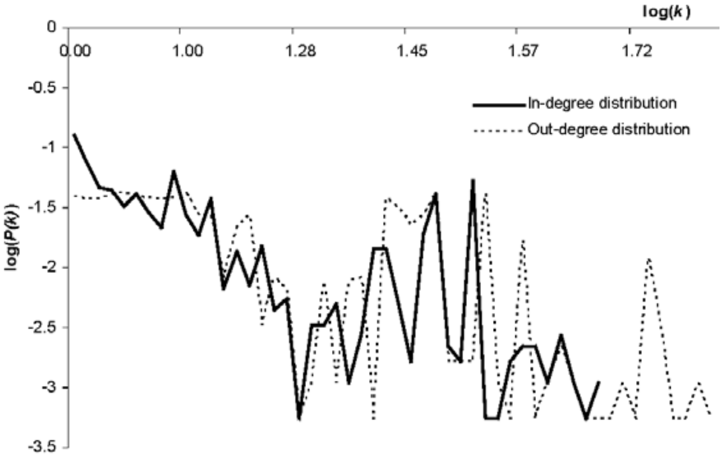
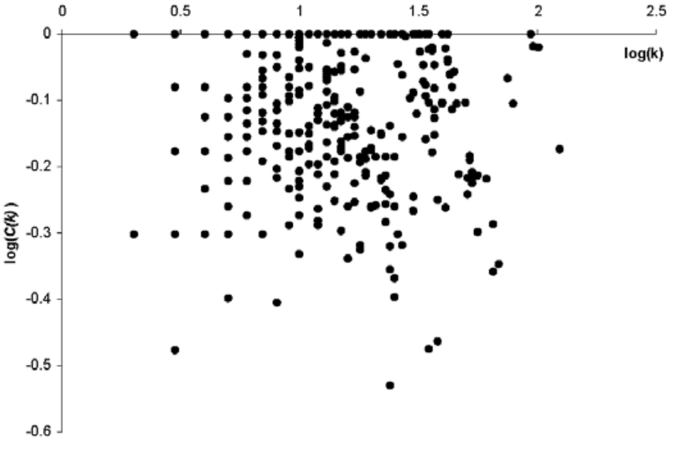


Figure 5 Clustering coefficient as a function of node degree for the yeast metabolic network. Here the clustering coefficient does not seem to follow the power law either, which suggests that there is no hierarchical modularity in our network



In order to demonstrate the use of context for visualisation with CDA projection algorithm, we retrieved a KEGG metabolic pathway with Gene Ontology (Ashburner et al., 2000) annotations for *S. cerevisiae* species. Figure 6 shows zoomed in result of that retrieval in the neighbourhood of the *tricarboxylic acid cycle* biological process, while the CDA projection of that graph is shown in Figure 7. In this projection the *tricarboxylic acid cycle* biological process is biased so that its incident edges have lower weights than the other edges of the graph. We can see that in this projection there are two main clusters. In one cluster there are the *tricarboxylic acid cycle* Gene Ontology term (Number 1) and its neighbour nodes. Therefore, we may conclude that in this metabolic pathway there is a group of enzymes and compounds that are strongly involved in the *tricarboxylic acid cycle* biological process and there is another group that is weakly involved in this process.

Figure 6 A zoom of a yeast metabolic pathway in the neighbourhood of *tricarboxylic acid (TCA) cycle* (GO:0006099). Proteins involved in the TCA cycle biological process are clustered near the TCA cycle Gene Ontology term

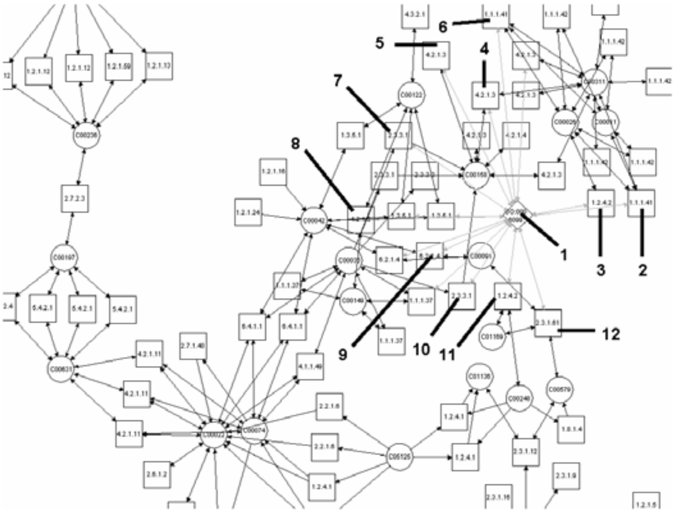
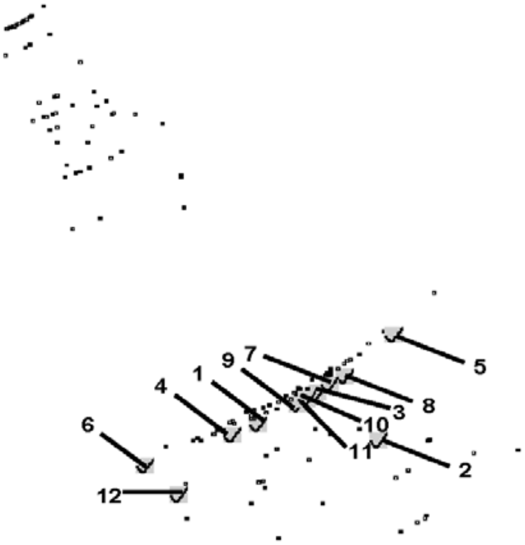


Figure 7 A Curvilinear Distance Analysis projection biasing *tricarboxylci acid cycle.* The projection was obtained by lowering the distance of all connected edges to TCA node (number 1) in the above graph



4.2 *Protein neighbourhood search as a context dependent annotation*

Assignment of protein function is a nontrivial task due to the fact that the same proteins may be involved in different biological processes, depending on the state of the biological system and protein localisation. Therefore, protein function is context dependent. Protein databases such as UniProt (Bairoch et al., 2005) contain information on protein function in text format. For example, PPAR gamma (UniProt id: P37231) is annotated as

> "Receptor that binds peroxisome proliferators such as hypolipidemic drugs and fatty acids. Once activated by a ligand, the receptor binds to a promoter element in the gene for acyl-CoA oxidase and activates its transcription. It therefore controls the peroxisomal beta-oxidation pathway of fatty acids. Key regulator of adipocyte differentiation and glucose homeostasis." (http://www.expasy.org/cgi-bin/niceprot.pl?P37231)

Such information may not be satisfactory if interested in the role of this protein in context of specific disease (PPARγ is known to be involved in a variety of diseases, such as diabetes, osteoporosis, and cancer), tissue localisation (PPAR gamma actually has two main isoforms, 1 and 2, of which PPAR gamma 1 is expressed in all tissues, while PPAR gamma 2 is mainly expressed in adipose tissue; we have been recently involved in the characterisation of the latter (Medina-Gomez et al., 2005), or relationship with a specific group of proteins. We have previously proposed the network based approach to annotate proteins in context dependent manner by using the 'protein neighbourhood search' (Gopalacharyulu et al., 2005), i.e., exploring the local relationships of proteins with other biological entities such as proteins, genes, biological processes etc.

As an illustration of the utility of the approach, we queried a select set of proteins related to regulation of energy homeostasis and to insulin signalling. The following human proteins have been queried:

- Peroxisome proliferator activated receptor gamma (PPARγ; UniProt id: P37231)
- Peroxisome proliferator activated receptor alpha (PPARα; UniProt id: Q07869)
- Peroxisome proliferator activated receptor gamma coactivator 1 alpha (PGC1α; UniProt id: Q9UBK2)
- Sterol regulatory element binding protein 2 (SREBP – 2; UniProt id: Q12772)
- Putative G protein-coupled receptor GPR40 (GPR40; O14842)
- Putative G protein-coupled receptor GPR41 (GPR41; O14843)
- Probable G protein-coupled receptor GPR43 (GPR43; O15552).

The resulting network is shown in Figure 8. Short descriptions of select entities in the network are presented in Table 2. While detailed study of the retrieved protein neighbourhood lies beyond the scope of this paper, we will show its use on one example. The entity numbered 10 (Protein arginine N-methyltransferase 2) does not have well assigned function. The UniProt resource lists the protein function as

> "Probably methylates the guanidino nitrogens of arginyl residues in some proteins. May play a role in transcriptional coactivation." (http://www.expasy.org/cgi-bin/niceprot.pl?P55345)

Our data suggests the protein is binding with PPARγ, and so may be related to regulation of energy homeostasis. This provides a hypothesis for designing new experiments to address the function of a protein that would have more likely escaped attention otherwise. The topic of transcriptional co-regulators involved in energy homeostasis is a topic of intense research in domains of diabetes and metabolic syndrome (Lin et al., 2005).

Figure 8 Query for proteins PPAR gamma, PPAR alpha, PGC1, SREBP 2, GPR40, GPR41, GPR43 in HUMANS. The numbered nodes are listed in Table 3. Grey lines are Gene Ontology relations, dark grey the regulatory networks, light grey the protein-protein interactions

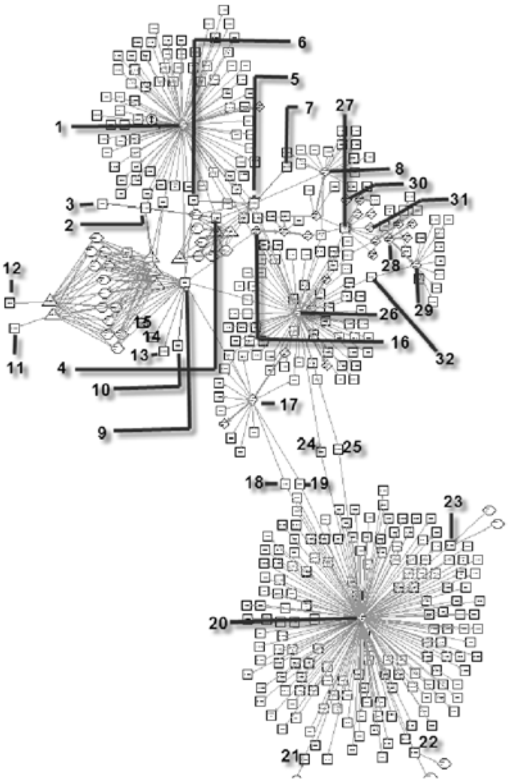


Table 2 Short description of select entities from the network shown in Figure 8

Label	*Name*	*ID* (*UniProt/GO accession*)	*Important interactions/associations* (*Identified by Labels 1–32*)
1	Lipid metabolism	GO:0006629	–
2*	Sterol regulatory element binding protein-2 (SREBP-2)	Q12772	3, 4 (MINT); 1 (GO)
3	Transcription factor SP1	P08047	2* (MINT)
4	Hepatocyte nuclear factor 4 aplha	P41235	2*(MINT); 1 (GO)
5*	Peroxisome proliferator activated receptor alpha	Q07869	5* (BIND); 6, 7 (MINT); 1, 8, 26 (GO)
6	Retinoic acid receptor RXR – alpha	P19793	5 *(MINT); 9* (TRANSFAC – interacting factor)
7	Nuclear receptor corepressor 2	Q9Y618	5* (MINT)
8	Fatty acid metabolism	GO:0006631	5* (GO)

Table 2 Short description of select entities from the network shown in Figure 8 (continued)

Label	*Name*	*ID* (*UniProt/GO accession*)	*Important interactions/associations* (*Identified by Labels 1–32*)
9*	Peroxisome proliferator activated receptor gamma	P37231	10 (BIND); 6,13,14,15 (TRANSFAC – interacting factors); 1,16,17,26 (GO)
10	Protein arginine N-methyltransferase 2	P55345; EC: 2.1.1	9* (BIND)
11	Nuclear factor of activated T-cells, cytoplasmic 4	Q14934	9* (TRANSFAC – transcription factor of)
12	CCAAT/enhancer binding protein alpha	P49715	9* (TRANSFAC – transcription factor of)
13	Nuclear factor of activated T-cells, cytoplasmic 1	O95644	9* (TRANSFAC – interacting factor)
14	Nuclear receptor coactivator 1	O00150; EC: 2.3.1.48	9* (TRANSFAC – interacting factor)
15	CREB-binding protein	Q92793; EC: 2.3.1.48	9* (TRNASFAC – interacting factor)
16	White fat cell differentiation	GO:0050872	9* (GO)
17	Response to nutrients	GO:0007584	9*, 18, 19 (GO)
18	Somatostatin precursor	P61278	17, 20 (GO)
19	Guanine nucleotide-binding protein G(i), alpha-2 subunit	P04899	17, 20 (GO)
20	G-protein coupled receptor protein signalling pathway	GO:0007186	18, 19, 21*, 22*, 23*, 24, 25 (GO)
21*	Putative G protein-coupled receptor GPR40	O14842	20 (GO)
22*	Putative G protein-coupled receptor GPR41	O14843	20 (GO)
23*	Probable G protein-coupled receptor GPR43	O15552	20 (GO)
24	Vasopressin V1a receptor	P37288	20, 26 (GO)
25	Melanin-concentrating hormone receptor 1	Q99705	20, 26 (GO)
26	Generation of precursor metabolites and energy	GO:0006091	5*, 9*, 24, 25, 32 (GO)
27*	Peroxisome proliferator activated receptor gamma coactivator 1 alpha	Q9UBK2	28, 30, 31 (GO)
28	Gluconeogenesis	GO:0006094	27*, 29 (GO)
29	Glucose metabolism	GO:0006006	32 (GO)
30	Positive regulation of histone acetylation	GO:0035066	27* (GO)
31	Thermoregulation	GO:0001659	27* (GO)
32	Insulin precursor	P01308	26, 29 (GO)

*Denotes an entity used in making the query for network construction.

Table 3 Short description of a few select entities from the network presented in Figure 6

Label	*Name/description*	*ID* (*UniProt/GO accession/EC number*)
1	tricarboxylic acid cycle	GO:0006099
2	alpha-4-beta-4 subunit of mitochondrial isocitrate dehydrogenase 1	P28834, 1.1.1.41
3	alpha-ketoglutarate dehydrogenase	P20967, 1.2.4.2
4, 5	Aconitase, mitochondrial	P19414, 4.2.1.3
6	NAD+-dependent isocitrate dehydrogenase	P28241, 1.1.1.41
7	Mitochondrial isoform of citrate synthase	P43635, 2.3.3.1
8	Fumarase; converts fumaric acid to L-malic acid in the TCA cycle. The GI molecule identifier below refers to the protein encoded by this gene	P08417, 4.2.1.2
9	alpha subunit of succinyl-CoA ligase (synthetase; ATP-forming), a mitochondrial enzyme of the TCA cycle	P53598, 6.2.1.4
10	citrate synthase. Nuclear encoded mitochondrial protein	P00890, 2.3.3.1
11	alpha-ketoglutarate dehydrogenase	P20967, 1.2.4.2
12	dihydrolipoyl transsuccinylase component of alpha-ketoglutarate dehydrogenase complex in mitochondria	P19262, 2.3.1.61

4.3 Type 1 Diabetes gene expression data

The network edges drawn in previous examples were based on existing knowledge resources such as pathways and ontologies. However, the network representation affords extension to other relationships, such as gene sequence similarity or co-regulation of molecules based on profiling experiments (or collection of multiple experiments). The former may be particularly useful when building metabolic models of species with unannotated genomes based on the existing metabolic models from well annotated species. The latter may be utilised to interpret the data obtained from molecular profiling experiments. For example, applications have been reported linking the gene co-expression obtained from micro-array experiments to functional modules in cancer cells (Segal et al., 2004). We have previously utilised the correlation network approach to integrate across metabolite, protein, and gene level experimental profile data (Oresic et al., 2004).

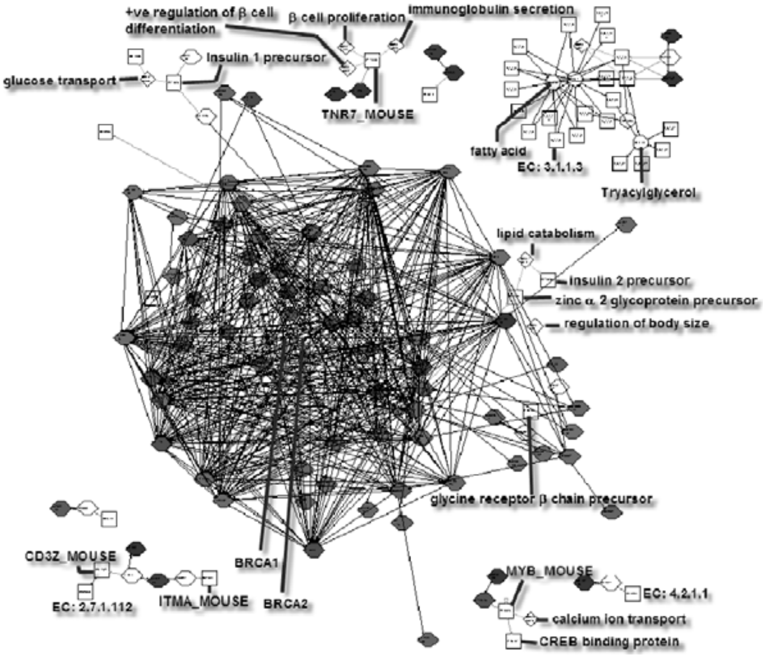
As an illustration of combining gene expression data with the existing pathways and ontologies, we utilised gene expression data from mouse congenic strains in a study related to Type 1 Diabetes (Eaves et al., 2002). We processed this data as explained below in order to construct the query. The resulting network is shown in Figure 9. Some relevant entities in network are indicated with their names. The gene expression data is incorporated as follows:

- Normalised dataset is downloaded from the NCI GEO database (www.ncbi.nlm.gov/geo). GEO accession number of the data is GDS10.
- Pearson correlation coefficients are calculated for every pair of genes.
- Based on distribution of correlation coefficients a cut-off correlation of 0.997 is set to select only highly correlated pairs (the cut-off can be varied as part of the exploratory analysis). One hundred and sixty six gene pairs pass this cut off.

- These gene pairs and their correlation values are defined as a relational table in Oracle database.
- We compared the Diabetic strain data with Non diabetic strain data from Spleen. The procedure for calculating the intensity ratios is explained below:
- The Average Intensity values (AI) contain negative values. Hence these values are shifted so that the least AI value becomes 1. AI values in all samples are shifted by a constant value of 49.
- Average of each group of samples is calculated.
- Ratio between average corresponding to diabetic samples is taken over average corresponding to non diabetic samples.
- These values are then visualised such that down regulated genes appear in green, up-regulated genes appear in red and expression level of each gene determines a colour between these two extremes.

The largest upregulated cluster is clearly related to lipid and glucose metabolism, but perhaps most curious finding being the upregulated BRCA1 and BRCA2 genes within this cluster. BRCA genes are associated with breast cancer, but are known to be highly expressed in spleen and associated with immune response. How these genes specifically relate to Type 1 Diabetes is unclear, and certainly this finding is worthy of further study. In another upregulated small cluster of genes we found association with beta-cell proliferation, which is a known response to increased rate of beta-cell apoptosis in Type 1 Diabetes.

Figure 9 Correlation network of gene expression data related to Type 1 Diabetes from Eaves et al. (2002)



5 Discussion

In this paper we introduced an approach and a system which affords integration, mining, and visualisation of systems biology data. Three examples were given in domains of network topology studies, context-dependent protein annotation, and integration of gene co-expression data with available pathway knowledge. It is evident that the studies of complex organisms such as mammals, for example in the context of drug discovery, generate datasets representing physiological processes at multiple spatial and temporal levels. This necessitates the data integration solutions that facilitate mining of such diverse data (Gopalacharyulu et al., 2005; Oresic et al., 2004; van der Greef and McBurney, 2005; Searls, 2005). Depending on availability of data, this may include building associations and dependencies across biological entities, either based on available knowledge such as ontologies or on mathematical models. As we have shown in this paper, these two approaches are not mutually exclusive.

Our integration approach is based on the premise that relationships between biological entities can be represented as a complex network. The information in such networks forms a basis for exploratory mining, as well as for development of predictive models. Distances between different nodes in an integrated network play a central role. In order to calculate distances, one first needs to define distance measures across heterogeneous types of information. We are taking a pragmatic approach by letting the user define the distances as a part of the query. This is reasonable since the distance basically defines the context of the questions posed by the user and allows biasing the similarity toward particular types of relationships, or towards a relationship in a specific context. Once the distance measure is specified, we can map the nodes of the graph into a lower dimensional space. We introduced and implemented three methods to perform such mappings: Sammon's mapping, CCA and CDA. As these mappings are approximate, there will be some distortion while doing the mapping. Therefore, in our opinion the exact form of distance measure is not a critical issue, as far as it underlines the relationships in the concept graph. In fact, selection of distance measure may reflect a subjective choice and as such will be subject to debate. It is ultimately the end result of mining that determines the utility of specific distance measure.

The three examples described in this paper demonstrate the utility of our approach. We show how the study of global network properties is facilitated using our approach. Similarly, the local properties of networks can be studied, as well as the properties of integrated networks (i.e., cross-talk between metabolism and cell signalling). Related to the second example, current annotation of proteins using e.g., Gene Ontology or UniProt do not take into account the complexity and context-dependency of protein function and interactions. We introduced a visual approach which enables context dependent interpretation. For example, in a query of six proteins related to energy homeostasis and insulin signalling we found a potential function for currently poorly annotated protein. We also extended the data integration framework to include experimental data. As a third example, we performed exploratory data analysis that linked clusters of gene expression profiles from spleen of NOD mouse model of Type 1 Diabetes to known interactions, regulatory pathways and ontologies related to the gene products within the clusters. While the 'pathway analysis' (Curtis et al., 2005) has already been widely utilised for analyses of gene expression data, our approach affords analysis across both physical interaction information (i.e., regulatory networks, protein-protein interactions, metabolic networks) as well as across known pathway annotations. As such it enables visual exploration of patterns found in data, facilitating to answer the first question any biologist is after when attempting to interpret high-dimensional micro-array data, i.e., what appears to be going on in the system based on the experimental evidence.

The pathway integration framework described in this paper is not limited only to the static biological pathways. Other models can be incorporated as well, as long as they are represented in the exchangeable schemas such as SBML or CellML. Our framework then affords further model refinement using interaction and ontology information from diverse sources. In addition, the metabolic models from well characterised species such as yeast (Förster et al., 2003) can be extended to less characterised related species. The data mining methods described in the paper are largely focused on integration across heterogeneous sources and mapping of complex networks into lower-dimensional space for the purpose of visualisation. What is needed is incorporation of more advanced data mining methods for statistical analysis and modelling of data. We believe the network framework opens new possibilities for analyses of complex heterogeneous life science data.

Currently our system is able to visualise data at molecular level. One of the remaining challenges would be to visualise multiple levels (Saraiya et al., 2005). This kind of approach would enable us to investigate how a small change at the molecular level affects the higher abstract level (e.g., tissue or organ level). Another appealing challenge would be to visualise biological networks in three dimensions (Changsu Lee and Park, 2002; Férey et al., 2005).

6 Conclusions

We presented an integrated database software system that enables retrieval and visualisation of biological relationships across heterogeneous data sources. We demonstrate the utility of our approach in three applications: full metabolic network retrieval with network topology study, exploration of properties and relationships of a specific set of proteins, and combined visualisation and exploration of gene expression data with related pathways and ontologies. We believe our approach facilitates discovery of novel or unexpected relationships, formulation of new hypotheses, design of experiments, data annotation, interpretation of new experimental data, and construction and validation of new network-based models of biological systems.

Acknowledgements

Matej Orešič was in part funded by Marie Curie International Reintegration Grant. The authors are thankful towards Teemu Kivioja, Laxman Yetukuri, and Jaakko Hollmén for helpful discussions during this work.

The authors Peddinti V. Gopalacharyulu and Erno Lindfors contributed equally to this work.

References

Arita, M. (2004) 'The metabolic world of Escherichia coli is not small', *Proc. Natl. Acad. Sci. USA*, Vol. 101, pp.1543–1547.

Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S. and Eppig, J. (2000) 'Gene ontology: tool for the unification of biology', *Nat. Genet*., Vol. 25, pp.25–29.

Bader, G.D., Betel, D. and Hogue, C.W.V. (2003) 'BIND: the biomolecular interaction network database', *Nucl. Acids Res*., Vol. 31, pp.248–250.

Bairoch, A., Apweiler, R., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N. and Yeh, L-S.L. (2005) 'The universal protein resource (UniProt)', *Nucl. Acids Res*., Vol. 33, pp.D154–159.

Barabasi, A-L. and Oltvai, Z.N. (2004) 'Network biology: understanding the cell's functional organization', *Nat. Rev. Genet*., Vol. 5, pp.101–113.

Bard, J.B.L. and Rhee, S.Y. (2004) 'Ontologies in biology: design, applications and future challenges', *Nat. Rev. Genet*., Vol. 5, pp.213–222.

Bodenreider, O. (2004) 'The unified medical language system (UMLS): integrating biomedical terminology', *Nucl. Acids Res*., Vol. 32, pp.D267–270.

Bollobás, B. (1998) *Modern Graph Theory*, Springer-Verlag, New York.

Changsu Lee, J.P. and Park, J.C. (2002) 'BiopathwayBuilder: nested 3D visualization system for complex molecular interactions', *Genome Informatics*, Vol. 13, pp.447, 448.

Chung, S.Y. and Wong, L. (1999) 'Kleisli: a new tool for data integration in biology', *Trends Biotechnol*., Vol. 17, pp.351–355.

Critchlow, T., Fidelis, K., Ganesh, M., Musick, R. and Slezak, T. (2000) 'DataFoundry: information management for scientific data', *IEEE Trans. Inf. Technol. Biomed*., Vol. 4, pp.52–57.

Curtis, K., Oresic, M. and Vidal-Puig, A. (2005) 'Pathways to analysis of microarray data', *Trends Biotechnol*, Vol. 8, pp.429–435.

Dahlquist, K.D., Karen Vranizan, N.S., Lawlor, S.C. and Conklin, B.R. (2002) 'GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways', *Nature Genetics*, Vol. 31, pp.19, 20.

Davidson, S.B., Overton, C.G., Tannen, V. and Wong, L. (1997) 'BioKleisli: a digital library for biomedical researchers', *Int. J. on Digital Libraries*, Vol. 1, pp.36–53.

Demartines, P. and Hérault, J. (1997) 'Curvilinear component analysis: a self-organizing neural network for nonlinear mapping of data sets', *IEEE Trans. Neur. Netw*., Vol. 8, pp.148–154.

Dijkstra, E. (1959) *Numerische Mathematik*, Vol. 1, pp.269–271.

Dorogovtsev, S.N. and Mendes, J.F.F. (2003) *Evolution of Networks from Biological Nets to the Internet and WWW*, Oxford University Press, Oxford, UK.

Eaves, I.A., Wicker, L.S., Ghandour, G., Lyons, P.A., Peterson, L.B., Todd, J.A. and Glynne, R.J. (2002) 'Combining mouse congenic strains and microarray gene expression analyses to study a complex trait: the NOD model of Type 1 diabetes', *Genome Research*, pp.232–243.

Etzold, T. and Argos, P. (1993) 'SRS – an indexing and retrieval tool for flat file data libraries', *CABIOS*, Vol. 9, pp.49–57.

Etzold, T., Ulyanov, A. and Argos, P. (1996) 'SRS: information retrieval system for molecular biology data banks', *Meth. Enzymology*, pp.114–128.

Férey, N., Hérisson, P.E.G.J. and Gherbi, R. (2005) 'Visual data mining of genomic databases by immersive graph-based exploration', *3rd International Conference on Computer Graphics and Interactive Techniques in Australasia and South East Asia (Dunedin, New Zealand, November 29–December 02, 2005), GRAPHITE '05*, ACM Press, New York, NY, pp.143–146.

Förster, J., Famili, I., Fu, P., Palsson, B.O. and Nielsen, J. (2003) 'Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network', *Genome Res*., Vol. 13, pp.244–253.

Gärdenfors, P. (2000) *Conceptual Spaces: The Geometry of Thought*, MIT Press, Cambridge, MA.

Gene Ontology Consortium (2000) 'Gene ontology: tool for the unification of biology', *Nature Genetics*, Vol. 25, pp.25–29.

Gopalacharyulu, P.V., Lindfors, E., Bounsaythip, C., Kivioja, T., Yetukuri, L., Hollmen, J. and Oresic, M. (2005) 'Data integration and visualization system for enabling conceptual biology', *Bioinformatics*, Vol. 21, pp.i177–185.

Gopalacharyulu, P.V., Lindfors, E., Bounsaythip, C., Wefelmeyer, W. and Oresic, M. (2004) 'Ontology based data integration and context-based mining for life sciences', *W3C Workshop on Semantic Web for Life Sciences*, Cambridge, MA.

Goto, S., Okuno, Y., Hattori, M., Nishioka, T. and Kanehisa, M. (2002) 'LIGAND: database of chemical compounds and reactions in biological pathways', *Nucl. Acids Res*., Vol. 30, pp.402–404.

Guimera, R. and Amaral, L.A.N. (2005) 'Functional cartography of complex metabolic networks', *Nature*, Vol. 433, pp.895–900.

Han, J.D., Bertin, N., Hao, T., Goldberg, D.S., Berriz, G.F., Zhang, L.V., Dupuy, D., Walhout, A.J., Cusick, M.E., Roth, F.P. and Vidal, M. (2004) 'Evidence for dynamically organized modularity in the yeast protein-protein interaction net-work', *Nature*, Vol. 430, pp.88–93.

Hass, L.M., Schwartz, P.M. and Kodali, P. (2001) 'DiscoveryLink: a system for integrated access to life science data sources', *IBM Systems Journal*, Vol. 40, pp.489–511.

Herman, I., Melancon, G. and Marshall, M.S. (2000) 'Graph visualization and navigation in information visualization: a survey', *IEEE CS Society*, Vol. 6, pp.24–43.

Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., Margalit, H., Armstrong, J., Bairoch, A., Cesareni, G., Sherman, D. and Apweiler, R. (2004) 'IntAct: an open source molecular interaction database', *Nucl. Acids Res*., Vol. 32, pp.D452–455.

Ideker, T., Galitski, T. and Hood, L. (2001) 'A new approach to decoding life: systems biology', *Annu. Rev. Genomics Hum. Genet.*, Vol. 2, pp.343–372.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. and Barabási, A-L. (2000) 'The large-scale organization of metabolic networks', *Nature*, Vol. 407, pp.651–654.

Jeong, H., Mason, S.P., Barabási, A-L. and Oltvai, Z.N. (2001) 'Lethality and centrality in protein networks', *Nature*, Vol. 411, pp.41, 42.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. and Hattori, M. (2004) 'The KEGG resource for deciphering the genome', *Nucl. Acids Res*., Vol. 32, pp.D277–280.

Kitano, H. (2002) 'Systems biology: a brief overview', *Science*, Vol. 295, pp.1662–1664.

Krull, M., Voss, N., Choi, C., Pistor, S., Potapov, A. and Wingender, E. (2003) 'TRANSPATH: an integrated database on signal transduction and a tool for array analysis', *Nucl. Acids Res*., Vol. 31, pp.97–100.

Lee, J.A., Lendasse, A. and Verleysen, M. (2004a) 'Nonlinear projection with curvilinear distances: isomap versus curvilinear distance analysis', *Neurocomputing*, Vol. 57, pp.49–76.

Lee, S.G., Hur, J.U. and Kim, Y.S. (2004b) 'A graph-theoretic modeling on GO space for biological interpretation of gene clusters', *Bioinformatics*, Vol. 20, pp.381–388.

Lin, J., Handschin, C. and Spiegelman, B.M. (2005) 'Metabolic control through the PGC-1 family of transcription coactivators', *Cell Metab*., Vol. 1, pp.361–370.

Luscombe, N.M., Madan Babu, M., Yu, H., Snyder, M., Teichmann, S.A. and Gerstein, M. (2004) *Genomic Analysis of Regulatory Network Dynamics Reveals Large Topological Changes*, Vol. 431, pp.308–312.

Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V., Kloos, D-U., Land, S., Lewicki-Potapov, B., Michael, H., Munch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S. and Wingender, E. (2003) 'TRANSFAC: transcriptional regulation, from patterns to profiles', *Nucl. Acids Res.*, Vol. 31, pp.374–378.

Medina-Gomez, G., Virtue, S., Lelliott, C., Boiani, R., Campbell, M., Christodoulides, C., Perrin, C., Jimenez-Linan, M., Blount, M., Dixon, J., Zahn, D., Thresher, R.R., Aparicio, S., Carlton, M., Colledge, W.H., Kettunen, M.I., Seppanen-Laakso, T., Sethi, J.K., O'Rahilly, S., Brindle, K., Cinti, S., Oresic, M., Burcelin, R. and Vidal-Puig, A. (2005) 'The link between nutritional status and insulin sensitivity is dependent on the adipocyte-specific Peroxisome Proliferator-Activated Receptor-{gamma}2 isoform', *Diabetes*, Vol. 54, pp.1706–1716.

Oresic, M., Clish, C.B., Davidov, E.J., Verheij, E., Vogels, J.T.W.E., Havekes, L.M., Neumann, E., Adourian, A., Naylor, S., van der Greef, J. and Plasterer, T. (2004) 'Phenotype characterization using integrated gene transcript, protein and metabolite profiling', *Appl. Bioinformatics*, Vol. 3, pp.205–217.

Ravasz, E. and Barabási, A-L. (2003) 'Hierarchical organization in complex networks', *Physical Review*, Vol. 67, pp.1–7.

Sammon Jr., J.W. (1969) 'A nonlinear mapping for data structure analysis', *IEEE Trans. Comp.*, Vol. C-18, pp.401–409.

Saraiya, P., North, C. and Duca, K. (2005) 'Visualization for biological pathways: requirements analysis, systems evaluation and research agenda', *IEEE Trans. Vis. Comput. Graph.*, Vol. 11, pp.443–456.

Searls, D.B. (2005) 'Data integration: challenges for drug discovery', *Nat. Rev. Drug Disc.*, Vol. 4, pp.45–48.

Segal, E., Friedman, N., Koller, D. and Regev, A. (2004) 'A module map showing conditional activity of expression modules in cancer', *Nat. Genetics*, Vol. 36, pp.1090–1098.

Shannon, P.M.A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) 'Cytoscape: a software environment for integrated models of biomolecular interaction networksar interactions', *Genome Research*, Vol. 3, pp.2498–2504.

van der Greef, J. and McBurney, R. (2005) 'Rescuing drug discovery: in vivo systems pathology and systems pharmacology', *Nat. Rev. Drug Disc.*, Vol. 4, pp.961–967.

Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M. and Cesareni, G. (2002) 'MINT: a molecular INTeraction database', *FEBS Lett.*, Vol. 513, pp.135–140.

Website

NCBI PubChem, http://pubchem.ncbi.nlm.nih.gov/.

Article III

Context dependent visualization of protein function

In: Juho Rousu, Samuel Kaski and Esko Ukkonen (eds.). Probabilistic Modeling and Machine Learning in Structural and Systems Biology. 2006, Tuusula, Finland, pp. 26–31.

Context dependent visualization of protein function

Peddinti V. Gopalacharyulu[1], Erno Lindfors[1], Catherine Bounsaythip[1],
and Matej Orešič[1]

[1] VTT Technical Research Centre of Finland, Tietotie 2,
FIN-02044 Espoo, Finland
{ext-peddinti.gopal, erno.lindfors, catherine.bounsaythip, matej.oresic}@vtt.fi

Abstract. Assignment of protein function is a nontrivial task due to the fact that the same proteins may be involved in different biological processes, depending on the state of the biological system and protein localization. Therefore, protein function is context dependent and textual annotations commonly utilized to describe protein function lack the flexibility to address such contextuality. We propose an alternative approach for protein annotation motivated by the conceptual space approach, which relies on context-driven mapping of complex relationships based on known protein interactions or ontologies and on experimental data into low-dimensional space. We utilize the curvilinear distance analysis to generate such mappings, and demonstrate the approach on a set of proteins involved in maintenance of energy homeostasis.

Keywords: Protein function, conceptual spaces, curvilinear distance analysis

1 Introduction

The wealth of information generated with modern life science technologies, combined with existing repositories of knowledge dispersed across numerous databases and literature, demand new solutions for management and integration of life science data. Biological systems are characterized by the complexity of interactions of their internal parts and also with the external environment. The protein repositories such as UniProt [1] describe the protein function in textual format. In such form it may be difficult if not impossible to express the protein function in a given context, therefore another layer of representation is necessary.

While biological ontologies such as Gene Ontology [2] attempt to unify part of our life science knowledge at the molecular level, the diversity of life science research and questions addressed inevitably lead to multiple and overlapping ontologies. In turn, these Ontologies need to be integrated and unified, a challenge addressed by the Semantic Web approaches. However, these approaches are mostly based on hard coded symbolic representations which are valid only if the context in which they were created is stable. Therefore, in the fast evolving knowledge in life science, such approaches lack flexibility, emergence and context sensitivity.

In this paper we propose a visual approach for context-dependent protein function characterization, motivated by P. Gärdenfors' paradigm of *conceptual spaces* [3]. The main idea behind conceptual spaces is that if we use a group of objects or “*clusters*” as references, they are much more reliable than single objects. In the conceptual spaces, *clusters* remain stable even when objects change their properties or when new objects come into existence or old ones disappear. Unlike in ontological structures, the name that is given to a cluster does not need to be taken as such by its sole semantic sense, but it is enriched by the set of qualities (called “*quality dimensions*”) of the cluster it represents. Therefore, naming convention is not a bottleneck as in Semantic Web approach.

In living systems the quality dimensions may correspond to different levels of biological organization, where the objects (e.g. molecules, cells, organs) and their quality dimension specific relationships can be described with certain geometric structures (in some cases they are *topological* or *orderings*). Therefore, with the aid of the dimensions, similarities between biological entities and concepts can easily be represented by the distance in a conceptual space.

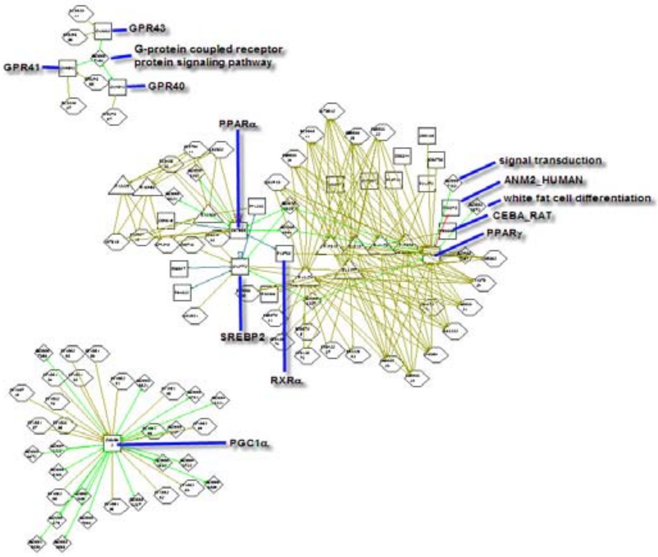

GPR43
G-protein coupled receptor protein signaling pathway
GPR41
GPR40
PPARα
signal transduction
ANM2_HUMAN
white fat cell differentiation
CEBA_RAT
PPARγ
SREBP2
RXRα
PGC1α

Fig. 1. Query for the protein neighborhood of PPARγ, PPARα, PGC1α, SREBP2, GPR40, GPR41, GPR43 human proteins, utilizing BIND [4], MINT [5], DIP [6], KEGG [7], Transfac [8] and Gene Onology [2] databases. Squares represent proteins, hexagons genes, triangles DNA binding sites, and diamonds GO terms.

2 Network representation

We represent networks as directed weighted graphs where biological entities are nodes connected via interactions or relationships between them [9]. In the context of protein function, a typical question utilizing network representation is about the *protein neighborhood*, i.e. what are the nearest nodes connected to a particular protein, with the edges being either direct interactions or ontology-defined relationships. Fig. 1 shows an example of a specific query for a set of human proteins related to maintenance of energy homeostasis and specific G-protein coupled receptors (GPCRs) that are not yet well characterized. The query for the nearest neighbor protein relationships revealed three distinct clusters, with all three GPCRs jointly in a separate cluster. While some of the well known relationships were revealed in the largest cluster, the results of the query have not facilitated characterization of poorly annotated proteins such as GPR40, GPR41, and GPR43.

3 Conceptual spaces and biological entities

In the network view illustrated above, a logical follow-up query would include extension of protein neighborhood search for the next-nearest neighbors or beyond. However, due to high-connectivity of biological entities such approach soon becomes visually prohibitive. As a pragmatic alternative, we define a distance metric for each type of relationship, and allow the user to assign weights to different types of relationships as part of the mining process [9]. The key problem then becomes how to efficiently map data to a lower dimensional space in order to be able to visualize them in a context-dependent manner. We implemented Curvilinear Distance Analysis (CDA) [10] in our system. Curvilinear distance depends not only on the two points between which the distance is measured but also on the other surrounding points. Intuitively, instead of computing straight distances between the points, the goal of curvilinear distance consists in computing distances along an object that can be, for example, curves on the surface or any set of points.

CDA maps the points in a higher dimensional space into a lower dimensional space by preserving the distances in the original space. It calculates curvilinear distances in the high dimensional input space by creating a graph out of centroids. After that it calculates distances between centroids using Dijkstra's shortest path algorithm [11]. CDA works by optimizing a criterion that explicitly measures the preservation of the pairwise distances:

$$E_{CDA} = \sum \left(\delta_{ij} - d_{ij}\right)^2 F(d_{ij}, \lambda),$$

where δ_{ij} is distance measured between points p_i and p_j in the high dimensional data space and d_{ij} is distance measured between the coordinates of the same two points in the projection space. The factor $F(d_{ij},\lambda)$ weighs the contribution of each pair of points in the criterion. F is implemented as the Heaviside unit step function:

$$\begin{aligned} F\left(d_{ij}, \lambda\right) = \theta\left(\lambda - d_{ij}\right) &= 0 \text{ if } \lambda - d_{ij} < 0 \\ &= 1 \text{ if } \lambda - d_{ij} \geq 0. \end{aligned}$$

Starting from the criterion, the derivation of the learning rule follows a similar scheme as for a stochastic gradient descent. Instead of moving one mapped point according to the position of all other ones, one point m_i is frozen while moving all others radially around it:

$$m_j \leftarrow m_j + \alpha F\left(d_{ij},\lambda\right)\left(\delta_{ij} - d_{ij}\right)\frac{m_j - m_i}{d_{ij}}\ ,$$

where α and λ are time decreasing learning rate and neighborhood radius, respectively.

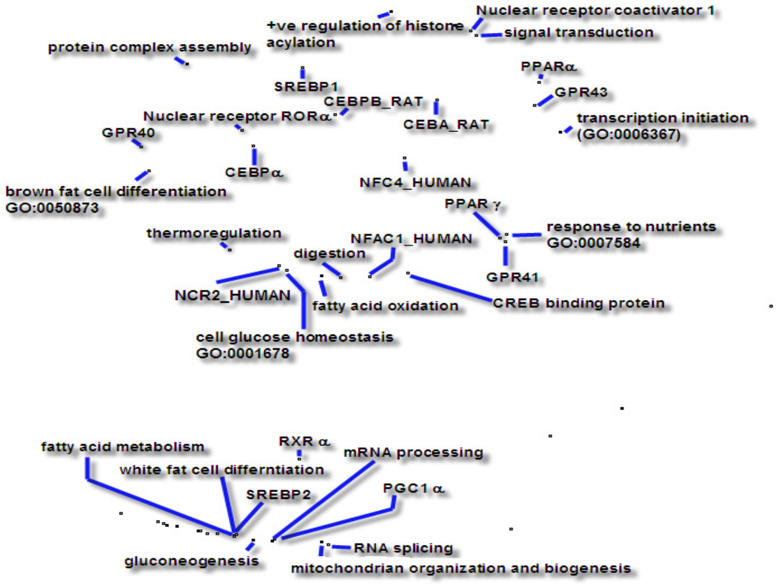


Fig. 2. Results of CDA mapping for the entities and databases listed in Fig. 1. All edge weights (unit costs) taken to be 1.

An application example of CDA mapping is shown in Fig. 2, querying for the same entities as in Fig. 1. While there are many interesting aspects of biology retrieved in the mapping, we focus here on PPARγ. PPARγ (UniProt id: P37231) is annotated in UniProt as "*Receptor that binds peroxisome proliferators such as hypolipidemic drugs and fatty acids. Once activated by a ligand, the receptor binds to a promoter element in the gene for acyl-CoA oxidase and activates its transcription. It therefore controls the peroxisomal beta-oxidation pathway of fatty acids. Key regulator of adipocyte differentiation and glucose homeostasis*". This is not a satisfactory explanation when searching for specific context, for example specific disease or relationship to specific GPCR. Our CDA projection revealed both PPARγ and GPR41 are closely associated with response to nutrients. Interestingly, this finding is supported by recent research [12], yet it cannot be revealed by searching any of the databases used individually.

Our approach is not limited only to pathway data and ontologies. Experimental data, such as gene expression or metabolomics experiments, can also be utilized to further define the context. In such cases the distance measure relating biological entities in the molecular profile space may correspond to the measure of co-expression (such as correlation coefficient) between different entities. Fig. 3 shows an example of CDA mapping based on a similar query as listed previously, but for the mouse proteins, and in the context of a specific gene expression dataset [13] from spleen tissue of NOD mouse. Curiously, several tumor suppressor genes such as BRCA1 associated with PPARγ, are found in this mapping. This finding deserves further attention. Only recently a link between a specific tumor suppressor (LKB1) and diabetes has been established [14], linking cancer and physiological control of metabolism.

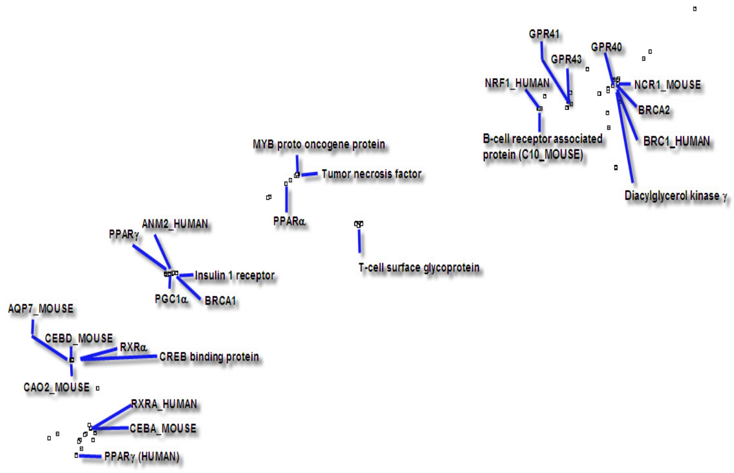


Fig. 3. Results of CDA mapping in context of Type 1 Diabetes for mouse proteins PPARγ, PPARα, PGC1α, GPR40, GPR41, GPR43 from databases listed above, plus the gene expression data from [13].

4 Conclusions

In this paper we introduced an approach aiming to facilitate mining of complex biological networks, ontologies, and high-dimensional molecular profile data. We focused specifically on context-dependent protein function assignment. The approach relies on network-based representation of biological entities, concepts, and their relationships, context-dependent assignment of distances between them, and nonlinear mapping into low-dimensional space to visualize distribution of concepts and entities in a specific context. Given the complexity of biological systems and fragmentation of biological knowledge, we believe our pragmatic approach is superior to more formal approaches such as based on Semantic Web technology in its flexibility and ability to extract potentially novel biological relationships leading to new hypotheses. For example, none of the surprising context-dependent functional relationships related to the PPARγ protein shown in this paper could be derived by mining Gene Ontology or other bioinformatics databases alone. Our approach also provides new opportunities for research of topological structures defined by complex biological relationships.

References

1. Apweiler, R., et al., *UniProt: the Universal Protein knowledgebase.* Nucleic Acids Research, 2004. **32**(Database issue): p. D115-119.
2. Ashburner, M., et al., *Gene ontology: tool for the unification of biology.* Nat. Genet., 2000. **25**: p. 25 - 29.
3. Gärdenfors, P., *Conceptual spaces: The geometry of thought.* 2000, Cambridge, MA: MIT Press.
4. Bader, G.D., D. Betel, and C.W.V. Hogue, *BIND: the Biomolecular Interaction Network Database.* Nucl. Acids Res., 2003. **31**(1): p. 248-250.
5. Zanzoni, A., et al., *MINT: a Molecular INTeraction database.* FEBS Lett., 2002. **513**(1): p. 135-140.
6. Salwinski, L., et al., *The Database of Interacting Proteins: 2004 update.* Nucl. Acids Res., 2004. **32**(suppl_1): p. D449-451.
7. Kanehisa, M., et al., *The KEGG resource for deciphering the genome.* Nucl. Acids Res., 2004. **32**(90001): p. D277-280.
8. Matys, V., et al., *TRANSFAC: transcriptional regulation, from patterns to profiles.* Nucl. Acids Res., 2003. **31**(1): p. 374-378.
9. Gopalacharyulu, P.V., et al., *Data integration and visualization system for enabling conceptual biology.* Bioinformatics, 2005. **21**(suppl_1): p. i177-185.
10. Lee, J.A., A. Lendasse, and M. Verleysen, *Nonlinear projection with curvilinear distances: Isomap versus curvilinear distance analysis.* Neurocomputing, 2004. **57**: p. 49-76.
11. Dijkstra, E., *A note on two problems in connexion with graphs.* Numerische Mathematik, 1959. **1**: p. 269-271.
12. Xiong, Y., et al., *Short-chain fatty acids stimulate leptin production in adipocytes through the G protein-coupled receptor GPR41.* PNAS, 2004. **101**(4): p. 1045-1050.
13. Eaves, I.A., et al., *Combining mouse congenic strains and microarray gene expression analyses to study a complex trait: The NOD model of Type 1 Diabetes.* Genome Research, 2002(12): p. 232-243.
14. Shaw, R.J., et al., *The Kinase LKB1 mediates glucose homeostasis in liver and therapeutic effects of metformin.* Science, 2005. **310**(5754): p. 1642-1646.

Article IV

Network-based representation of biological data for enabling context-based mining

In: Catherine Bounsaythip, Jaakko Hollmen,
Samuel Kaski and Matej Orešič (eds.). Proceedings
of KRBIO´05, International Symposium on
Knowledge Representation in Bioinformatics.
2005, Espoo, Finland, Helsinki University of
Technology, Laboratory of Computer and
Information Science, pp. 1–6.
Reprinted with permission from the publisher.

NETWORK-BASED REPRESENTATION OF BIOLOGICAL DATA FOR ENABLING CONTEXT-BASED MINING

Catherine Bounsaythip[1], *Erno Lindfors*[1], *Peddinti V. Gopalacharyulu*[1], *Jaakko Hollmén*[2], *Matej Orešič*[1]

[1]VTT Biotechnology,
P.O. Box 1500, Espoo, FI-02044 VTT, Finland, *name.surname*@vtt.fi, ext-Gopal.Peddinti@vtt.fi,
[2]Helsinki University of Technology,
Laboratory of Computer and Information Science, P.O. Box 5400, Espoo, FIN-02015 HUT,
Finland, Jaakko.Hollmen@hut.fi

ABSTRACT

Biological phenomena are usually described by relational model of interactions and dependencies between different entities. Therefore, a network-based knowledge representation of biological knowledge seems to be an obvious choice. In this paper, we propose such a representation when integrating data from heterogeneous life science data sources, including information extracted from biomedical literature. We show that such a representation enables explanatory analysis in a context dependent manner. The context is enabled by a judicious assignment of weights on the quality dimensions. Analysis of clusters of nodes and links in the context of underlying biological questions may provide emergence of new concepts and understanding. Results are obtained with our *megNet* software, an integrative platform based on a multi-tier architecture using a native XML database.

1. INTRODUCTION

The primary goal of knowledge representation is to enable computer to assist humans in analyzing complex forms of data to discover useful information. This has resulted in a wide range of techniques and tools. How to represent knowledge depends largely on the way reasoning can be done with that knowledge. For example, early works have been mainly focused on logic-based representation. Recently, techniques combining machine learning, pattern recognition, statistics, and artificial intelligence have been employed. Although these are well-developed disciplines, their applications in life science have been limited [1][2][3].

Biology is a data rich discipline. The problem is that this source of knowledge is stored in a large number of different data sources which need to be mined in parallel. Integrating all this information and its efficient mining is a challenge with huge application potential [4][5]. Moreover, each database may have its own interface that users may not have time to adequately learn to use them efficiently. A tool which can integrate the mining as well as visualization of heterogeneous life science data would therefore open new possibilities for the exploration of biological knowledge and possibly lead to novel discoveries.

As biological systems are characterized by the complexity of interactions of their internal parts and also with the external environment, integrating such interacting information may result in a large connected graph with nodes and edges of heterogeneous types. This makes such information hard to visualize, and sophisticated methods have been developed for analyzing such complex networks [6][7][8][9]. The most important aspect in visualizing high-dimensional data in a lower dimensional space is how to preserve the proximity relationships. In practice, it is very difficult if not impossible to project hundreds of dimensional data to a smaller dimensional space (2 or 3 dimensions) in such a way that all similarity relationships are preserved. Therefore, in order to enable effective reasoning, the challenge is to find the best compromises by choosing which kinds of relationships to visualize and with what type of metrics to use in order to ensure the trustworthiness of the visualized data [10].

Another way to enable effective reasoning is to limit the scope of deliberations to a small context associated with the domains under consideration. This may be approached by assigning weights to the “quality dimensions” [11] under consideration (gene-centric, tissue-centric, compound-centric, disease-centric etc.)

The above criteria have been our motivations to develop an integrated visualization tool, *megNet*, that uses topological analysis of complex networks to visualize query results in a single interface. It also enables context-based information display from our integrated database system (see [12]).

This paper discusses the representation and visualization aspects of our integration platform. It is organized as follows: Section 2 discusses about the network representation and clustering methods, including the notion of distance and context. Section 3 gives examples of visualizing a protein-protein interaction network.

2. BIOLOGICAL NETWORKS

With the growing trend towards systems biology, integrated biological networks contain many different types of entities and attributes arising from a growing number of disparate data sources, including literature databases. These databases have been created by different scientific communities, for different purposes, and covered different aspects. All that led to a high level of structural and semantic heterogeneity. The structural and semantic integration aspects of these databases have been reported in our previous papers [12][13]. Here we will focus on the retrieval and visualization of these heterogeneous data. We are mainly interested in the data from the following databases:

- Protein-protein interaction databases: *BIND* [14], *DIP* [15], and *MINT* [16].
- Biochemical pathways database: *KEGG* [17].
- *TransFac* is a database on DNA binding elements and their transcription factors [18].
- *TransPath*, an extension of *TransFac*, contains signal transduction pathways that regulate the activity of transcriptional factors in different species [19].
- *GeneOntology* (GO) is a database of three structured controlled vocabularies that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner [20].

The first step after retrieving all the massive information from databases is to build the network. The objects in network are then clustered based on some similarity measure for the display. The definition of the similarity measure is thus a crucial step.

2.1. Network representation

The graph representation contains nodes and edges [21][22]. The nodes include various kinds of molecules, e.g., proteins, compounds, genes, mRNAs etc. For example, in the case of protein-protein interaction network, we would relate the neighboring proteins by searching all the possible pathways among them, including their regulating genes. The generated nodes and edges show the proteins and their interactions, respectively.

Our biological network is presented as a directed weighted graph where biological entities are nodes that are connected to each other through edges which are interactions between the entities. The shape of the nodes will be coded differently depending on the type of an entity. The edges can be directed or undirected depending on the nature of the interactions (Figure 1).

A metabolic network consists of *reactions*. In one reaction there are *substrates*, *products* and at least one *enzyme* that catalyzes the reaction. The substrates, products and enzymes are presented as nodes. The substrates and products are presented as circles and the enzymes are presented as squares. Since some reactions are reversible and other reactions are irreversible, directed edges are used to distinguish the direction of a reaction. But in a protein-protein interaction network, interactions between the proteins are represented with undirected edges, because the interaction is mutual.

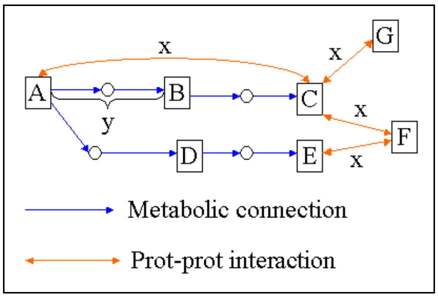


Figure 1: Example of our integrated network representation used. The distance between the entities A and B, is the same as for B to A. If there is not any path between two nodes, we assume that the distance between them is infinity.

The shortest path length between each entity is obtained by using Tom Sawyer Java analysis toolkit (Tom Sawyer, Inc.). The distances between each entity in both directions are calculated, based on the cost of connection types. In Figure 1, the cost of a metabolic interval is denoted by y, and x is the cost of a protein-protein interaction. By changing these cost parameters we can investigate how protein-protein interactions affect the structure of metabolic pathways.

2.2. Clustering of biological networks

The molecular entities of the cell form a very complicated and dynamic interacting system. Yet, it has been demonstrated that this complex interactions shared some common network properties, e.g. the presence of single modularity networks [24][25][26]. However, the presence of the modularity in highly integrated biological networks is not self-evident as it lacks quantitative support [24]. There is thus a need for tools to identify the modularity of a biological network and to identify the modules and their relationships. Clustering is a mathematical method which allows the identification of key connectivity patterns of a network. The most common methods used when investigating the structure of complex networks are hierarchical clustering tree, Kohonen's Self-Organizing Maps (SOM) [28], and Sammon's mapping [29][30].

All clustering algorithms share the basic steps:

1. *Compute distance matrix;*
2. *Find closest pair of clusters;*
3. *Update distance matrix.*

First, the distance matrix must be computed. The distance matrix define distances from one entity to the other entities. The distance matrix from the graph represented in Figure 1 is:

$$D=\begin{bmatrix} 0 & y & \min(x,2y) & y & 2y & \min(x,2y)+x \\ 3y+x & 0 & y & \text{inf} & y+2x & y+x \\ x & x+y & 0 & x+y & \min(x+2y,2x) & x \\ y+3x & 2y+2x & y+2x & 0 & y & y+x \\ 3x & 3x+y & 2x & 3x+y & 0 & x \\ 2x & 2x+y & x & 2x+y & x & 0 \end{bmatrix}$$

If the purpose of the distance calculations is to investigate the structure of metabolic pathways, the distance matrix would not take into account metabolites and other proteins that do not belong to the metabolic pathway (e.g. entities F and G in Figure 1).

After the distance matrix has been obtained, we can apply clustering algorithm which will merge objects in the same cluster based on the self-similarity. The self-similarity of a group of elements is defined as the average pairwise similarity between the elements. One may also choose other criteria such that the pair of clusters maximizes the minimum similarity or minimize the maximum similarity.

Since the purpose of the distance matrix is to describe the proximity of the entities, the more similar distance vectors are, the closer are corresponding biological entities. In our current implementation, we use the Sammon's mapping algorithm to investigate the similarities of the distance vectors.

2.2.1. *Similarity measure*

For integrated network where entities are of complex nature, evaluating similarity is not a trivial task. While distances within the molecular networks can be intuitively set to the length of the shortest path between the molecules, distance measure is less obvious for relationships such as in ontologies. It was shown that GeneOntology can be represented as a graph, and the distance measures based on the shortest path to a common ancestor were already studied [31]. In the case of gene expression network which consists only of genes, the similarity measure is based on the gene expression level.

The challenge is to combine topology metrics and the quantitative information from the data. For instance, one can combine the gene expression level and the topology of the network in the same distance function such as in [32]: $d = f(\delta_{exp} + \delta_{net})$.

Given a set of data points x_i, let us note by $d(x_i, x_j)$ being the distance between two data points.

If we consider the gene expression level G_{ik} as a log-ratio gene expression of gene $g_{i,}$, the distance function could be based on the Pearson correlation coefficient:

$$\rho_{exp}(g_i, g_j) = \frac{1}{N}\sum_k \left(\frac{G_{ik} - \mu_i}{\sigma_i}\right)\left(\frac{G_{jk} - \mu_j}{\sigma_j}\right)$$

with μ_i and σ_i are mean and standard deviation of the transformed time series data of g_i .

The correlation coefficient is then converted to a distance function as a degree of dissimilarity with: $\delta_{exp}(g_i, g_j) = 1 - \rho(g_i, g_j)$. We obtain the combined distance function:

$$d(x_i, x_j) = 1 - 0.5 \times (\delta_{exp}(g_i, g_j) + \delta_{net}(v_i, v_j))$$

The network distance function could be based on the shortest path and the weighting function based on the degree of vertices.

It is supposed that this combined function may lead to increased stability of clustering solution when the gene expression levels support the relations in the networks and vice versa [32].

In our current implementation, gene expression databases are not yet fully operational for integrated mining.

2.2.2. *Data projection and non-linear mapping*

The main purpose of data projection is to transform a high dimensional data to a lower dimensional space in order to be able to visualize them. The Kohonen's self-organizing map (SOM) [28] is one popular method. But the delicate part of SOM is that the user needs to set control parameters carefully that may require sometimes *a priori* knowledge about the data. We have chosen the Sammon's mapping [29] as is easier to implement.

Like the SOM algorithm, the basic idea of the Sammon's mapping algorithm is to arrange all the data points on a 2-dimensional plane in such a way, that the distances between the data points in this output plane resemble the distances in vector space as defined by some metric as faithfully as possible. Unlike SOM algorithm, the Sammon's mapping algorithm tries to preserve internal distances in the input data that the human eye can easily detect. The structure of the input data is thus preserved through the mapping.

More formally, let d_{ij} be an element of a distance matrix D in input space, let o_i be the image of the data item x_j in the 2-dimensional output space. With O we denote the distance matrix containing the pairwise distances between images as measured by the Euclidean vector norm $\|o_i - o_j\|$. The goal is to place the o_i in such a way that the distance matrix O resembles as closely as possible matrix D, i.e. to optimize an error function E by following an iterative gradient-descent process:

$$E = \frac{1}{\sum_i \sum_{j>i} d_{ij}} \sum_i \sum_{j>i} \frac{(d_{ij} - \|o_i - o_j\|^2}{d_{ij}}$$

The resulting visualization depicts clusters in input space as groups of data points mapped close to each other in the output plane. Thus, the inherent structure of the original network can be derived from the structure detected in the 2-dimensional visualization.

2.3. Context

When a representation includes several domains, one must take into account the context in which what domains appear more or less important (or *salient*) [9].

Including context can be achieved by assigning *weights* to each domain. The relative weight of a domain will depend on the context.

2.3.1. *Weights as context dependent variables*

In the previous section, the distance function could be weighted as follows:

$$D_{ij} = \sum_{k=1}^{n} w_k d_{ijk}$$

The weights w_k can be seen as *context-dependent* variables that represent the relative degree of salience for each dimension. This aspect has been used in the subspace clustering algorithms which assume that cluster may exist in different subspaces of different sizes. For example, in the COSA algorithm [33], the weights are assigned to each dimension for each instance, not each cluster. Higher weights are assigned to those dimensions that have a smaller dispersion within the *k-nearest* group. The neighborhoods for each instance become iteratively enriched with instances belonging to its own cluster. The dimension weights are refined as the dimensions relevant to a cluster receive larger weights. This process enables some dimensions to emerge by different the clustering criteria. However, in the COSA algorithm, the number of dimensions to be included in a cluster cannot be set directly by the user, it is done through a parameter λ, which controls the incentive for clustering on more dimensions.

This COSA distance was shown to be more powerful than traditional Euclidean distance.

Therefore, the choice of the similarity measure can affect greatly the quality of the visualization in the projection space. When we change dimension in the visualization, the degree of similarity between two data points changes with the salience of the dimensions of the objects. This aspect was investigated in [9].

It must be noticed also that the knowledge and interest of the user may influence the "salience weights" as it is assumed that people can have different "perspectives". Therefore it is important that the user has also the possibility to influence this parameter in the visualization tool.

2.3.2. *The effect of context in knowledge discovery*

With the explosion of information resources on the Web, ontologies have been extensively developed to facilitate the understanding, sharing, re-use and integration of knowledge through the construction of an explicit domain model. In life science, the efforts in building ontologies across domains still have many challenges to go through [34][35]. Gene Ontology (GO) is the only ontology that has been extensively used in bioinformatics [36][37]. However, GO seems to be more a taxonomy rather than a well-formed ontological structure that would enable traditional rule-based reasoning [38]. Another drawback of GO and other Ontologies in general, is their static structure and thus, when used as a structure for reasoning, they can only produce *monotonic* inference. Such a mode of reasoning may hinder or possibly even prevent the discovery and exploration of new possibilities [39].

While in a context-based reasoning, the conceptualization associated to the "cluster" that has emerged from the context, is *non-static*. For example, when we interpret clusters obtained from gene expression data, we must take into account the context of underlying biological models e.g., from which tissue and what was environmental history which has led to that state.

3. EXAMPLES

In this section we would like to give an example of network clustering of data retrieved from metabolic pathways and protein-protein interaction databases. As an example, we create a network based on the KEGG metabolic pathway from the query: "Glycolysis / Gluconeogenesis, Pentose phosphate and Citrate cycle pathways", for *S. cerevisiae* (Figure 2). The enzymes are then enriched with protein-protein interaction (MINT, DIP).The query results are shown in Figure 3. We can see from the Sammon's mapping that there are two main clusters in these pathways, a strongly connected cluster and sparsely connected cluster (Figure 3). Sparsely connected proteins are highlighted with gray marks, which appear to be mostly located at the border of the graph. Based on the concept of hierarchical modularity, we may conclude that the proteins of the strongly connected cluster are in higher hierarchy level than those of the sparsely connected cluster.

Another example of search is performed for protein-protein interaction with the set of proteins {P41940, O15305, P29952} which are involved in the glycosylation and mannosylation pathways in S. *cerevisiae*, referenced in GeneOntology Biological process "GDP-mannose biosynthesis" with GO:0009298. Results are shown in Figure 5. Clustering examples with different contexts (different weight assignments) are given in Figure 6 and Figure 7. In Figure 6, all the edges have equal weights. We can see that he neighborhood of GO:0009298 consist of proteins C05345 and C00275, which denote that in this context, they have stronger connection to GO:0009298. In Figure 7, the neighbors of GO:0009298 have larger weights, this has resulted in the clustering of proteins of the query set {P41940, O15305, P29952}.

We can "experiment" with the weight assignment for different context and notice that relative proximity of nodes changes. This might suggest new hypotheses that these entities might be involved in the same process or pathways reflected by the context.

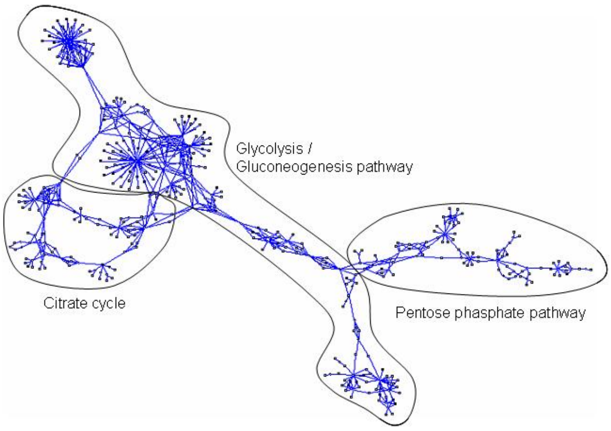


Figure 2: KEGG metabolic pathways for "Glycolysis / Gluconeogenesis , Pentose phosphate and Citrate cycle pathways.

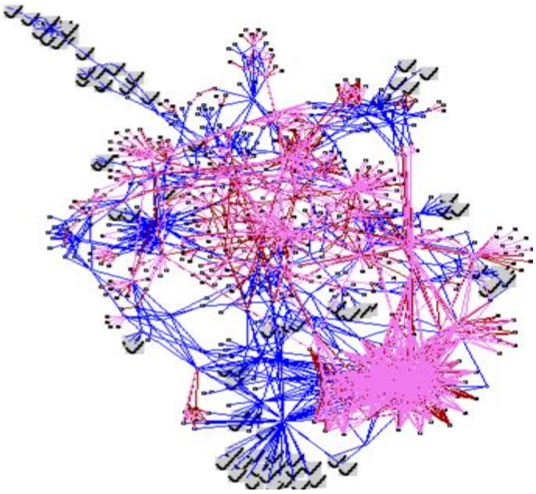


Figure 3: Metabolic pathway (KEGG) enriched with protein-protein interactions from MINT and DIP databases for “Glycolysis / Gluconeogenesis, Pentose phosphate and Citrate cycle pathways,. The proteins loosely connected are highlighted with gray marks.

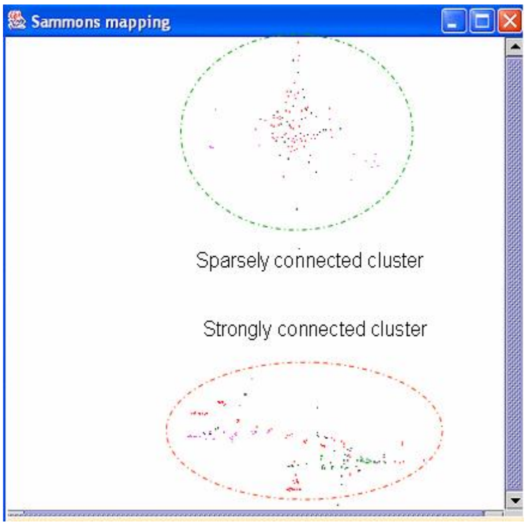


Figure 4: Clusters from Sammon’s mapping of the previous graph. Two main clusters emerged, one strongly connected and one loosely connected.

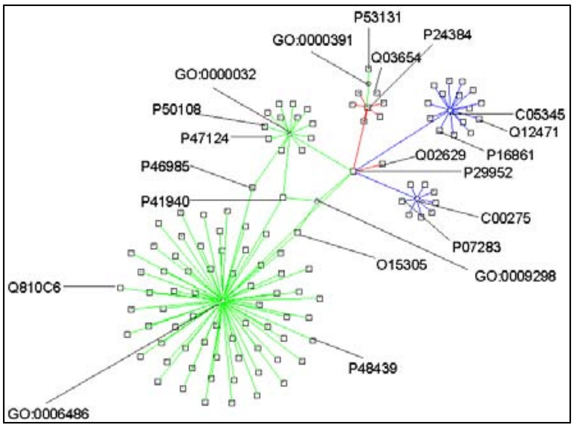


Figure 5: Search result of pathway query for mannose synthesis GO:0009298.

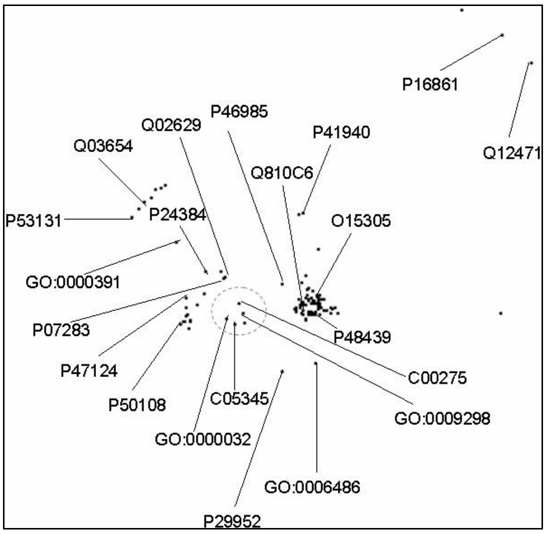


Figure 6: Sammon’s mapping of the previous network for ”Context 1: Every edge has equal weight”.

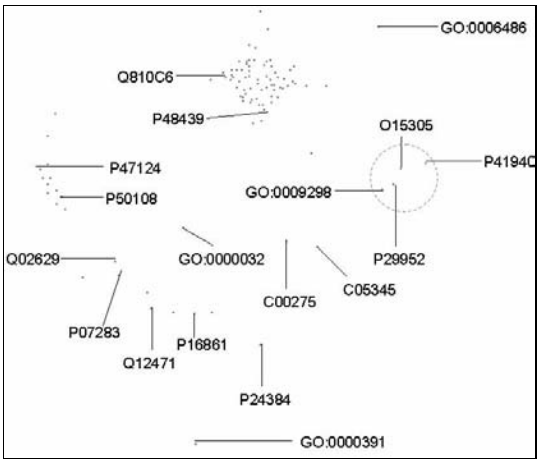


Figure 7: The Sammon’s mapping for “Context 2: The neighborhood edges of GO:0009298 have higher weights than the other edges”.

4. CONCLUSION

In this paper we have discussed about the heterogeneity of biological data and resources and existing methodologies to analyze those data. We introduced our approach to represent integrated biological data for enabling visual exploratory analysis. At the current phase, we have implemented the Sammon’s mapping clustering with a distance function that incorporates the notion of context, which can be controlled by the user. Our experiments have shown that the Sammon’s mapping algorithm is not very suitable for a large number of input vectors. Therefore, in our biological networks consisting of a large number of nodes, clustering time is rather long. Second, one cannot always rely totally on the output by the Sammon’s mapping clustering due to the trustworthiness of distance function. Therefore, it is up to the user to look for insight and experiment with the dimension salience to see if it makes any sense and always reconnect to the original hypothesis and background knowledge.

5. REFERENCES

[1] F. Capra, *The Web of Life*, Harper Collins, London, 1997.

[2] D. B. Kell, S. G. Oliver, " Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era", *Bioessays*; 26(1), pp. 99-105, 2004.

[3] F. Katagiri. "Attacking Complex Problems with the Power of Systems Biology", *Plant Physiology*, Vol. 132, pp. 417–419, 2003.

[4] D. B. Searls, "Data integration: challenges for drug discovery". *Nature Reviews Drug Disc*., 4, pp. 45-48, 2005.

[5] R. B. Stoughton, S. H. Friend, "How molecular profiling could revolutionize drug discovery", *Nature Rev. Drug Disc.*, Vol. 4, pp. 345-350, 2005.

[6] H. Jeong, B. Tombo, R. Albert, Z.N. Oltvai, A.-L. Barabási, "The Large-Scale Organization of Metabolic Networks", *Nature*, vol. 407, p. 651, 2000.

[7] M. E. J. Newman "The structure and function of complex networks", *SIAM Review*, 45(2), pp. 167- 256, 2003.

[8] A.-L. Barabási and Z. N. Oltvai, "Network Biology: Understanding the Cells' Functional Organization", *Nature Reviews Genetics*, vol. 5, pp. 101-114, Feb. 2004

[9] J. A. Papin, B. O. Palsson, "Topological analysis of mass-balanced signaling networks: a framework to obtain network properties including crosstalk". *J. Theor. Biol.*, 227, pp. 283-297, 2004.

[10] S. Kaski, J. Nikkilä, M. Oja, J. Venna, P. Törönen and E. Castrén, "Trustworthiness and metrics in visualizing similarity of gene expression", *BMC Bioinformatics*, pp. 4-48, 2003.

[11] P. Gärdenfors, *Conceptual spaces: The geometry of thought*, MIT Press, Cambridge, MA, 2000.

[12] P.V. Gopalacharyulu, E. Lindfors, C. Bounsaythip, T. Kivioja, L. Yetukuri, J. Hollmén, and M. Orešič, "Data integration and visualization system for enabling conceptual biology", *Proc. of International conference on Intelligent Systems for Molecular Biology* (ISMB 2005), Detroit, MI, USA, June 25-29, 2005.

[13] P. V. Gopalacharyulu, E. Lindfors, C. Bounsaythip, W. Wefelmeyer & M. Orešič, "Ontology based data integration and context-based mining for life sciences", *Proc. W3C Workshop on Semantic Web for Life Sciences*, Cambridge, MA, USA, 2004.

[14] G. D. Bader , D. Betel, C. W. V.Hogue, "BIND: the Biomolecular Interaction Network Database", *Nucl. Acids Res.*, 31, pp. 248-250, 2003.

[15] The DIP database, http://dip.doe-mbi.ucla.edu/

[16] A. Zanzoni, L. Montecchi-Palazzi, M. Quondam, G. Ausiello, M. Helmer-Citterich, G. Cesareni, "MINT: a Molecular INTeraction database", *FEBS Lett.*, 513, pp.135-140, 2002.

[17] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, M. Hattori, "The KEGG resource for deciphering the genome", *Nucl. Acids Res.*, 32, pp. 277-280, 2004.

[18] V. Matys, E. Fricke, R. Geffers, E. Gossling, *et al.* "TRANSFAC: transcriptional regulation, from patterns to profiles", *Nucl. Acids Res.*, vol. 31, pp. 374-378, 2003.

[19] M. Krull, N. Voss, C. Choi, S. Pistor, A. Potapov, E. Wingender, "TRANSPATH: an integrated database on signal transduction and a tool for array analysis", *Nucl. Acids Res.*, 31, pp. 97-100, 2003.

[20] M. Ashburner, C. Ball, J. Blake, D. Botstein, H. Butler, J. Cherry, A. Davis, K. Dolinski, S. Dwight, J. Eppig, "Gene ontology: tool for the unification of biology", *Nature Genetics*, vol. 25, pp. 25–29, 2000.

[21] B. Bollobás, *Modern Graph Theory*, Graduate Texts in Mathematics, vol. 184, Springer, New York, 1998.

[22] R. Diestel, *Graph Theory*, Graduate Texts in Mathematics, vol. 173, Springer, New York, 1997.

[23] S.N. Dorogovtsev, J.F.F. Mendes, *Evolution of Networks From Biological Nets to the Internet and WWW*, Oxford University Press, Oxford, UK, 2003.

[24] A.-L. Barabási, Z. N. Oltvai, "Network Biology: Understanding the Cells' Functional Organization", *Nature Reviews Genetics*, vol. 5, pp. 101-113, 2004.

[25] J. D. Han, N. Bertin, T. Hao, D.S. Goldberg, G.F. Berriz, L. V. Zhang, D. Dupuy, A.J. Walhout, M.E. Cusick, F.P. Roth, M. Vidal," Evidence for dynamically organized modularity in the yeast protein-protein interaction network", *Nature*, vol. 430, pp. 88-93, 2004.

[26] R. Guimera, L. A. Nunes Amaral, "Functional cartography of complex metabolic networks", *Nature*, vol. 433, pp. 895-900, 2005.

[27] E. Ravasz, A.-L. Barabási, "Hierarchical organization in complex networks", *Physical Review*, vol. 67, pp. 026112, pp. 1-7, 2003.

[28] T. Kohonen, *Self-Organizing Maps*, Springer Verlag, 2001.

[29] J. W. Sammon Jr., "A nonlinear mapping for data structure analysis". *IEEE Trans. Comp.*, C-18, 401-409, 1969.

[30] F. Azuaje, H. Wang, A. Chesneau, "Non-linear mapping for explanatory data anlysis in functional genomics", *BMC Bioinformatics*, pp. 6-13, 2005.

[31] S. G. Lee, J. U. Hur, Y. S. Kim, "A graph-theoretic modeling on GO space for biological interpretation of gene clusters". *Bioinformatics*, vol. 20, pp. 381-388, 2004.

[32] D. Hanisch, A. Zien, R. Zimmer, T. Lengauer, "Co-clustering of biological networks and gene expression data", *Bioinformatics*, Vol. 18, pp. 145-154, 2002.

[33] J. F. Friedman, J. J. Meulman, "Clustering objects on subsets of variables". *Journal of the Royal Statistical Society*, Series B, 4, pp. 815-849, 2004.

[34] R. Stevens, C. Wroe, P. Lord, C. Goble, "Ontologies in bioinformatics". *Handbook on Ontologies in Information Systems*, pp. 635-657, Springer, 2003.

[35] J. L. Bard, S. Y. Rhee, "Ontologies in biology: design, applications and future challenges", *Nature Review Genetics*, vol. 5(3), pp. 213-22, 2004.

[36] M. A. Harris et al. "The Gene Ontology (GO) database and informatics resource", *Nucleic Acids Res.* vol. 32 Database issue, pp. 258-261, 2004.

[37] E. Camon, M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N. Harte, R. Lopez, R. Ap-weiler. "The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology". *Nucl. Ac-ids Res.*, vol. 32, pp. 262-266, 2004.

[38] B. Smith, J. Williams, S. Schulze-Kremer, "The Ontology of the Gene Ontology", *in Proc. of the Annual Symposium of the American Medical Informatics Association*, Washington DC, Nov. 2003.

[39] C. Catton, D. Shotton, "The use of Named Graphs to enable ontology evolution", *W3C Workshop on the Semantic Web for Life Sciences*, Cambridge, MA, USA, 2004.

• • • VTT PUBLICATIONS 732 DATA INGEGRATION, PATHWAY ANALYSIS AND MINING FOR...

Systems biology aims to facilitate understanding of cellular behaviour in terms of interactions among cellular components and bio molecules, and the dynamics resulting from the interactions. Systems biology studies typically involve perturbation of a biological system, application of high-throughput *omics* technologies to measure biological components and interactions among them, and finally integration of the data into a biological network to understand the system's behaviour. Thus, data integration is a necessary step in systems biology. Data integration forms a basis for data mining, visualisation, as well as study of dynamics. This thesis presents methods for integration and mining of heterogeneous biological data, and for study of dynamic changes in response to system perturbations using these data.

ISBN 978-951-38-7385-1 (soft back ed.)
ISSN 1235-0621 (soft back ed.)

ISBN 978-951-38-7386-8 (URL: http://www.vtt.fi/publications/index.jsp)
ISSN 1455-0849 (URL: http://www.vtt.fi/publications/index.jsp)